UNIVERSITY OF HELSINKI

Then shall I know fully

Relative frequencies of part-of-speech n-grams in native and translated Finnish literary prose

Matias Tamminen
Pro gradu thesis
Master's Programme in Translation and Interpreting
Faculty of Arts
University of Helsinki
May 2018

Tiedekunta/Osasto – Fakultet/Sektion – Faculty Humanistinen tiedekunta	Laitos – Institution – Department Kielten osasto	
Tekijä – Författare – Author Matias Tamminen		
Työn nimi – Arbetets titel – Title Then shall I know fully: Relative frequencies of part-of-speech n-grams in native and translated Finnish literary prose		
Oppiaine – Läroämne – Subject Käännös- ja tulkkausviestintä, englanti		
Työn laji – Arbetets art – Level Pro gradu -tutkielma	Aika – Datum – Month and year toukokuu 2018	Sivumäärä – Sidoantal – Number of pages 58 s. + suomenkielinen lyhennelmä 12 s.

Tiivistelmä – Referat – Abstract

Tutkin pro gradu -tutkielmassani käännetyn ja supisuomalaisen kaunokirjallisen proosan sanaluokka-n-grammien eli *n* sanaa pitkien sanaluokkaketjujen suhteellisten frekvenssien eroja. Tarkoituksenani on etsiä käännetylle suomelle tyypillisiä syntaktisia piirteitä ja katsoa, voiko löydösten perusteella sanoa jotakin aiemmin esitettyjen käännösuniversaalihypoteesien paikkansapitävyydestä. Metodini on jatkokehitetty Borinin ja Prützin (2001) artikkelissaan käyttämästä metodista.

Aineistonani on kaksi korpusta, *Käännössuomen korpus* ja *Englantilaisen ja amerikkalaisen kirjallisuuden klassikoita Kersti Juvan suomentamina, englanti–suomi-rinnakkaiskorpus*, joista ensin mainittu sisältää supisuomalaista kaunokirjallista proosaa ja useista eri lähtökielistä suomennettua kaunokirjallista proosaa ja jälkimmäinen Kersti Juvan suomentamia romaaneja ja niiden lähtötekstit.

Tulosten perusteella voidaan sanoa, että tilanteissa, joissa lähtösysteemissä on vain yksi tapa ilmaista jokin tietty asia mutta kohdesysteemissä lähtösysteemin tavan lisäksi muita tapoja, lähtösysteemin tapa yliedustuu ja muut tavat aliedustuvat käännöksissä. Tämä on sekä *interferenssiä* eli lähtösysteemin vaikutusta kohdetekstiin että *kohdekielen uniikkiainesten aliedustumista* (Tirkkonen-Condit 2004).

Näiden universaalien lisäksi myös *konventionaalisuus*-käännösuniversaalihypoteesi saa aineistosta tukea. Konventionaalisuus näkyy niin interjektioiden frekvensseissä kuin mahdollisesti sanajärjestyksessäkin.

Tulevaisuudessa metodia voitaisiin kehittää käyttämällä sanaluokkakategorioiden sijasta dependenssisuhdekategorioita n-grammien tarkasteltavana ominaisuutena. Tutkimuksessa havaittujen ilmiöiden tunnistamisella voisi olla sovelluksia niin automaattisessa käännösevaluoinnissa, konekääntämisessä kuin kääntäjänkoulutuksessakin.

Avainsanat – Nyckelord – Keywords
deskriptiivinen käännöstiede, korpuspohjainen käännöstiede, käännösuniversaalit, käännöskieli

Säilytyspaikka – Förvaringställe – Where deposited
Helsingin yliopiston kirjasto – Helda/ E-thesis (opinnäytteet), ethesis.helsinki.fi

Muita tietoja – Övriga uppgifter – Additional information
Suomenkielisen lyhennelmän otsikko: Kerran tietoni on täydellistä: sanaluokka-n-grammien suhteelliset frekvenssit käännetyssä ja supisuomalaisessa kaunokirjallisessa proosassa

Table of Contents

List of Tables

List of Figures

1 Introduction

Digital comparable corpora became available for translation research around the turn of the millennium (Laviosa-Braithwaite 1996: 37–40, 53–113). One of the foremost research questions was the hypothesis of translation universals. By translation universals, we understand systematic and language independent tendencies of translated language as compared to either the source texts or comparable texts native to the target system. (Baker 1993: 242–245.)

The early studies were mainly lexical, in the absence of automatic annotation tools for higher levels of grammar. Borin and Prütz (2001) proposed using part-of-speech tagging to gain evidence of syntactic phenomena, such as word order. The paper remained inconclusive, and the theme was not pursued further.

Recently, in the wake of advances in computational linguistics, the quality of automatic grammar annotation is surging up. More sophisticated digital corpus work platforms are accessible to translation researchers. This motivates another look at the questions left open by the pioneers.

In this thesis, I shall study the differences in relative frequencies of part-of-speech n-grams (i.e. strings of part-of-speech tagged words that are *n* words long, such as the trigram NOUN VERB SCONJ) between corpus material of native Finnish literary prose and translated Finnish literary prose. By doing this, I try to reach syntactical (i.e. sentence structural) features of translated Finnish. I want to see whether the possible differences are predictable in the light of various proposed translation universals (see Chapter 3.2).

This research is motivated by the possible applications of the knowledge extracted by the means of the method I shall use. If structural manifestations of, e.g., source language *interference* on the target text can be identified, the knowledge can be used in machine learning to train computers to identify such manifestations. As machine translation (MT) fluency approaches human quality, the automatic evaluation of MT also needs to move closer to human translation evaluation. This lends new interest to the study of translations as compared to native language use.

At the same time, translation corpora and corpus annotation tools have become more efficient and accessible to human translation students than when corpus translation studies became in vogue. Knowledge about typical *interference* patterns could be used in translator training and (human) translation evaluation.

This work uses current corpus tools to see if translation theoretically relevant differences can be spotted between human translations and native texts using n-grams on part-of-speech annotated corpora. The results of the pilot may contribute to the search for more sensitive indicators of variation between translations and other text genres. Translation studies and translation technology are thus becoming directly relevant to one another.

First, in Chapter Two, I introduce previous research on the topic. Then, in Chapter Three, I go through the field of descriptive translation studies and the tradition of corpus-based translation studies. After that, in Chapter Four, I introduce the various corpora that make up my research material and the method I am using. In Chapter Five, I present and analyze the results. Lastly, in Chapter Six, I draw my conclusions.

2 Previous research

The topics of translated language, translation universals, and part-of-speech distribution have been previously researched in many ways. In this chapter, I go through some of these studies.

2.1 Lexis of translated Finnish

There has been some research in testing the proposed translation universals with corpora containing translated Finnish. Mauranen (2004) looks into lexis by comparing word frequencies of native and translated Finnish fiction in order to find out if and how source language interference manifests itself. The results show that translated language is different from native language and that the translations from different source languages also differ from each other, showing source language influence. (Mauranen 2004: 76–78.)

Various specific lexical studies of translated Finnish have also been carried out. Mauranen and Tiittula (2004) look into the first person singular pronoun in

translations from and to English and German and find out that the first person singular pronoun *minä/mä* is more frequent in Finnish translated from English and German than in native Finnish, *I* is more frequent in native English than English translated from Finnish, and that the first person singular pronoun *ich* is more frequent in German translated from Finnish than in native German. Their findings also show that the use of the first person singular pronoun decreases in translation from English and German into Finnish and increases in translation from Finnish into English and German, i.e. the translations are between the source and target systems in their use of the first person singular pronoun. (Mauranen and Tiittula 2004: 40–43, 66, 68.)

2.2 Syntax of translated Finnish

In addition to lexical features, the syntax of translated Finnish has also been studied. Puurtinen (2005) studies features that distinguish translated children's literature and native children's fiction. She has found out that non-finite structures are more frequent in translated text, colloquial language more frequent in original language, some connectors are more frequent in translated language, some in original language, and translators of children's literature tend to avoid repetition of verbs in reporting clauses. She concludes that the frequency of non-finite structures might run counter to the simplification and explicitation universals, while the scarcity of colloquial language is in harmony with the conventionalization universal and the avoidance of repetition is a proposed universal in itself (see Chapter 3.2). (Puurtinen 2005: 213–221.)

Eskola (2005) compares the frequencies of certain syntactic non-finite structures, namely referative, final, and temporal constructions as well as participial attributes and post-modifying comitatives, between native Finnish literature and translated Finnish literature from two different source languages: English and Russian. According to her, some of those structures are more frequent in translated and some in native text and the differences can be attributed to the existence/non-existence of source language stimuli. (Eskola 2005: 240.)

Inspired by Puurtinen and Eskola, Pulla (2011) also delves into the syntax of translated Finnish by looking at the frequency of temporal structures in non-fiction, namely economical texts. Her findings show that temporal structures are slightly

more frequent in translated economic texts than native ones, which is in line with earlier studies (Pulla 2011: 51). She concludes that the translation universal hypothesis of *simplification* (see Chapter 3.2) is not supported by her findings (Pulla 2011: 55–56).

2.3 Part-of-speech distribution

In addition to translated Finnish and translation universals, part-of-speech frequencies have also been studied. Heikkinen, Lehtinen, and Lounela (2001) offer some basic statistics over the Finnish language in general. Hudson (1994) observes that there are some regularities in part-of-speech frequencies over genre and even language boundaries.

There have not been many instances where the part-of-speech frequencies have been used in testing the translation universals. In an innovative paper, Borin and Prütz (2001) compare frequencies of part-of-speech n-grams between native and translated English. Although their article is regrettably brief (they do not really analyze their results), it lays out the foundations for this research, for I follow their method to a large extent.

3 Theoretical background

In this chapter, I introduce the field of descriptive translation studies, the concept of translation universals, and the history of utilizing corpora to study translated language.

3.1 From source and prescription to target and description

In Translation Studies, the focus of research has historically been on the source text. Translations were studied by comparing them to their respective source texts, and translation choices were evaluated on basis of whether they were equivalent to their source texts. (Baker 1993: 233–235.) This state of affairs began to change in the 1990s, as calls were made to shift the focus from the relation between the source and target texts to the relation between the target text and the target system (Baker 1993: 236) and from prescription (how things ought to be) to description (how things actually are) (Toury 1995: 1–5).

The shift from prescription to description was largely due to the growing dissatisfaction towards the lack of systematic and sound methodology in the field of Translation Studies (Toury 1980: 81, Baker 1993: 240). The first ideas were put forward by James Holmes, as he included a sub-field named "Descriptive Translation Studies" alongside "Theoretical Translation Studies" as a "pure" branch of Translation Studies, this "pure" branch being in opposition to "Applied Translation Studies", in his map of the field of Translation Studies, a first tentative step in meta-structuring the science (Toury 1995: 9–10, Holmes 1972/2004: 184). The idea of descriptive translation studies is to study "translations and translation practices" as "*observational* facts" that exist "irrespective of any prior theoretical consideration" and to test hypotheses supplied by translation theories (Toury 1980: 80). Toury went on to develop Holmes' map further: the empirical findings of descriptive studies should be extrapolated into general theories, and the general theories should then be tested out using descriptive methods, the descriptive and theoretical branches thus feeding each other. The various practitioners of the applied branch, e.g. translation teachers and translation critics, could then draw their own conclusions about good translation practices, but this relation between the "pure" and "applied" branches should, according to Toury, be "unilateral and indirect". (Laviosa-Braithwaite 1996: 24–25, Toury 1995: 15–19.) This three-way process is not unlike, say, physics, where field work and extrapolation reciprocally formed what we know as laws of physics, from which then engineers have made their own conclusions about what, e.g., bridges should be like.

The shift from source text orientation to target text orientation was intermingled with the pursuit of description. The target text orientation was built upon the work of Itamar Even-Zohar on polysystem theory (Baker 1993: 237–238). Even-Zohar (1979: 292) has stated that "standard language cannot be accounted for without the *non*-standard varieties; – – translated literature is not disconnected from 'original' literature". This validated the investigation of translated language and its relation to the corresponding original language (Baker 1993: 238). Also contributing to target text orientation was Frawley, who concluded that translations form a so-called "third code", a system different from both source and (native) target systems (Frawley 1984: 168–169).

3.2 Translation Universals

This shift towards target-orientation gave birth to the idea of *translation universals*, "patterns which are specific to translated texts" (Baker 1993: 242). Baker proposes a short list of candidates for this universal status. These are *explicitation* (translations being more explicit than the source texts and the comparable texts in the target system), *simplification/disambiguation* (translations being syntactically simpler and less ambiguous than their source texts) (whether these are the same thing or possibly two different universals is debatable), *conventionality* (unconventional or ungrammatical units being replaced with conventional ones in translation), *avoidance of repetition*, *exaggeration of target language features*, and *untypical frequencies* (some features being less or more prevalent in translations than in the source texts or comparable original target system texts). (Baker 1993: 243–245.)

The list of proposed translation universals has been amended by additional candidates since then. These candidates include *under-representation of unique items of the target language* (lower frequency of such items that lack an "obvious linguistic stimulus – – in the source text") (Tirkkonen-Condit 2004: 177–178) (this universal candidate being the polar opposite of *exaggeration of target language features*), *untypical collocations* (compared to comparable target system texts) (Mauranen 2000: 120), and *interference* (the general notion of the source text/system influencing the translations) (Toury 1995: 274–275).

Of these additions, the notion of *interference* as a universal (or a law as Toury [1995: 274] calls it) is interesting, because when introducing the concept of universals, Baker stated, explicitly and twice, that universals are patterns that "are not the result of interference" (Baker 1993: 242, 243). This conflict is addressed by Mauranen (2004: 66). According to her, one problem is already in the term *interference* itself: it is often used in a neutral manner to refer to the influence of the source language on the translation (this is the definition I follow), but also sometimes contrasted with *transfer*, in which case *transfer* means "positive" source language influence and *interference* "negative" source language influence (Mauranen 2004: 67).

The conclusion that I subscribe to is that a general, language pair and direction independent interference tendency is a possible translation universal in itself (see

Mauranen 2004: 79). When a proposed universal is not language pair and direction independent but retraceable to a specific source language feature, it is not a translation universal but a phenomenon that Eskola (2004: 85) calls *local translation law* (contrasted with *universal translation law* aka *translation universal*). It is also to be noted that the proposed universal *under-representation of unique items of the target language* is basically said to exist because of a lack of interference (Tirkkonen-Condit 2004: 183) and needs *interference* as a complementary universal in order to exist. Moreover, what I mean by "influence of the source language on the translation" in the above definition of *interference* is that the translation is closer to the source system than comparable native target language texts. If knowledge about the source language makes the translator to hypercorrect themselves and make the translation fall further from the source system than the comparable texts, the phenomenon would fall under *untypical frequencies* or *exaggeration of target language features*.

As seen in the list of proposed universals above, some universals contrast the translations to their source texts and others to comparable texts in the target system. The first to explicitly write about this difference was Andrew Chesterman (2004: 39–40), who calls the former group *S-universals* and the latter *T-universals*. In this study, I mainly concentrate on *T-universals* (and on *interference*, about which typological guesses can be made), because my method derived from Borin and Prütz (2001) is designed for comparing differences between translated texts and comparable native texts in the same language (even though Borin and Prütz, quite questionably, apply the method over language boundaries). Due to this, the source language side of the parallel corpus I use (CEALen, see Chapter 4.3.1) remains rather under-utilized.

However, one could apply the method of comparing the relative frequencies of part-of-speech n-grams to the source–target pair by comparing the native frequencies of both languages and then looking into whether the translations fall closer to the native source language frequencies than the comparable native target language texts. This way, at least possible manifestations of *interference* could be extracted.

3.3 Corpora as tools of descriptive translation studies

As the field of descriptive translation studies provides us with the empirical way of thinking and the translation universals provide us with hypotheses to be tested, what remains to be described is the methodology. What is clearly needed is corpora, i.e. collections of texts that are stored in electronic format and selected pertaining to selected criteria (Olohan 2004: 1).

Baker (1993: 245) calls for studies that use corpora as tools to capturing patterns between translations and comparable original texts in the same language and by doing this, either forming new translation universal hypotheses or confirming/disproving existing ones. She also offers (Baker 1995: 230–235) a typology of corpora that are used to carry out research in Translation Studies:

- *Parallel Corpora* contain translations and their respective source texts aligned to each other
- *Multilingual Corpora* are sets of traditional monolingual corpora in multiple languages
- *Comparable Corpora* contain translations and comparable texts of the same genre that are originally produced in the target language of the translations.

Comparable corpora did not exist at the time of Baker's articles (Baker 1993: 245, Baker 1995: 234) but have come into existence since then, as researchers (including Baker herself [1996: 178] together with her student Sara Laviosa-Braithwaite [1996: 53–84]) have followed her advice. I, too, use a comparable corpus, the *Corpus of Translated Finnish*, which, although around 20 years old, is the largest of its kind in Finnish. The other corpus I use, *Classics of English and American Literature translated by Kersti Juva, English–Finnish Parallel Corpus* (Juva 2018), is a parallel corpus according to this typology.

4 Material and method

In this chapter, I shall present my material and the methods I use. First, I explain the idea of using part-of-speech strings as a proxy for syntactical structures. Second, I go

through the corpora that I use and how I prepare them. Third, I present the different ways of finding meaningful differences in the frequencies of parts-of-speech.

4.1 Part-of-speech n-grams as a proxy for syntactical structures

We are looking for syntactical features of language. However, Finnish and other agglutinative languages are notoriously difficult for computers to syntactically annotate. For this reason, I use parts of speech as a proxy for syntactical categories such as dependency relations, because dependencies often have a certain part-of-speech category they prefer, such as subject preferring a noun (phrase). If I were to conduct this research in a language pair easier for parsers (such as Borin and Prütz' Swedish and English), I would use dependency relations instead, but, for now, the part-of-speech strings or *n-grams* have to suffice.

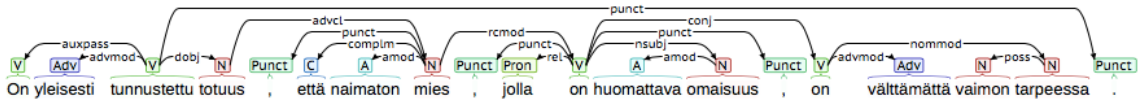

Figure 1: Dependency relations and part-of-speech tags[1]

An n-gram is a string of consecutive units where *n* is the number of units. A part-of-speech n-gram is an n-gram where the units are words and the examined attribute the part of speech of each word.

Here	is	an	example	.
ADV	VERB	DET	NOUN	PUNCT

Table 1: An example sentence on the concept of n-grams

In the example sentence above, there are

[1] The picture is taken from a previous version of CEAL (Juva 2017, see Chapter 4.3.1).

- five unigrams (ADV, VERB, DET, NOUN, PUNCT)
- four bigrams (ADV VERB, VERB DET, DET NOUN, NOUN PUNCT)
- three trigrams (ADV VERB DET, VERB DET NOUN, DET NOUN PUNCT)
- two 4-grams (ADV VERB DET NOUN, VERB DET NOUN PUNCT)
- one 5-gram (ADV VERB DET NOUN PUNCT).

As illustrated in the example, punctuation marks count as words and punctuation is a part-of-speech category.

4.2 Managing digital corpus workflows with Mylly

In order to get at the relative frequencies of n-grams, the raw data needs to be prepared. If (and in my case when) the data is not distributed with tokenization and annotation, those steps need to be taken. Then, the n-grams need to be calculated, counted, and normalized. This all calls for a software solution to managing corpus preparation.

For this preparation, I use the software *Mylly* [the Mill] by the Language Bank of Finland (Kielipankki, a). Mylly is a version of *Chipster* (see Chipster) specifically made for language analysis. Chipster is a modular computational platform originally developed for bioinformatics (Chipster). On this platform, one may run different tools on any material one has imported to the session and create workflows in which the output of one operation can act as an input for another. Mylly has a graphical user interface, which makes it quite easy to use. Thus, with Mylly, anyone can analyze language material quantitatively.

To describe a typical Mylly workflow, I explain the process of calculating relative frequencies of n-grams from a corpus with the help of Figure 2 below.

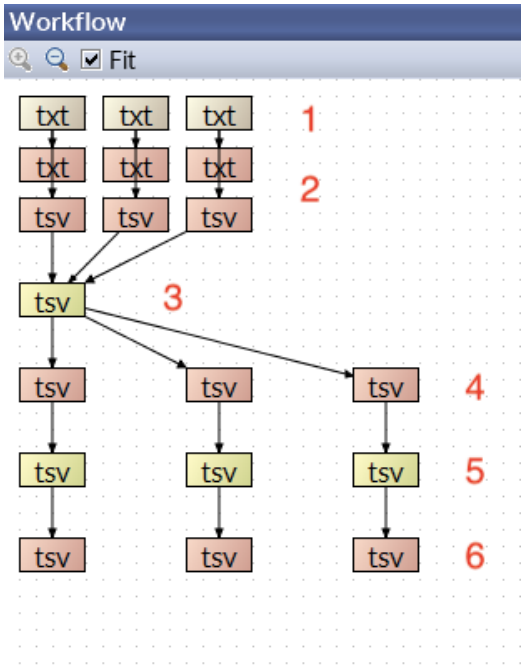

Figure 2: The process of calculating relative frequencies of n-grams

Every rectangle in the workflow window represents one file. The arrows between files show which files are calculated from which.

The three beige files at number 1 are the books of this example corpus in plaintext. The first step in this process is the annotation of the books with a parser. This produces the red txt and tsv files at number 2. Here, the computer has marked the part-of-speech category each word belongs to next to the word in question.

At number 3, there is the summation of the three annotated books. This step is simply making three files into one.

The step at number 4 is the calculation of n-grams – in this case, from left to right, unigrams, bigrams, and trigrams. The frequencies are not yet calculated at this point – the software simply marks each n-gram and produces the red tsv files.

The step at number 5 is the calculation of the absolute frequencies of the n-grams. The computer counts how many instances of each n-gram there are in the data.

The last step is to calculate the relative frequencies, i.e. the percentage values, from the absolute ones. This produces the red tsv files at number 6.

4.3 The corpora

In this study, I utilize two different corpora that are divided into smaller sub-corpora. I shall present the corpora and the sub-divisions in this sub-chapter.

4.3.1 Classics of English and American Literature

My main corpus of translated Finnish literary prose is the corpus *Classics of English and American Literature in translated by Kersti Juva, English–Finnish Parallel Corpus* (henceforth CEAL) (Juva 2018)[2]. It contains Kersti Juva's Finnish translations of Jane Austen's novel *Pride and Prejudice*, Henry James' novel *Washington Square*, and Charles Dickens' novel *Bleak House* as well as the aligned source texts in English (Juva 2018). This makes the corpus a parallel corpus in Baker's (1995: 230–235) typology. The Finnish sub-corpus (henceforth CEALfi) of the target texts contains 502,062 word-tokens. The English sub-corpus (henceforth CEALen) of the source texts contains 657,986 word-tokens.

I chose this corpus in addition to the *Corpus of Translated Finnish* introduced below, because this corpus is one of the few parallel corpora of translated Finnish (i.e. it contains not only the translations but also their source texts aligned) and because Kersti Juva is perhaps the best-known translator of literary prose in Finland and her translations are widely regarded as being of high quality (Juva and Hartikainen 2014). The corpus shall be readily available from Kielipankki (Kielipankki, b), although one has to apply for a data permission.

4.3.2 Corpus of Translated Finnish

The other corpus that I am using is the *Corpus of Translated Finnish* (henceforth CTF). The corpus is a large collection of hand-selected texts in different genres, in both native Finnish and translated Finnish from multiple source languages. (The Corpus of Translated Finnish.) This makes the corpus a comparable corpus in Baker's (1995: 230–235) typology. I utilize the sub-corpora of native Finnish literary prose (henceforth SKA), translated Finnish literary prose from English (henceforth KKAen), translated Finnish literary prose from Russian (henceforth KKAru), and

[2] I thank Kersti Juva for giving me access to this version of the corpus before it was officially published.

translated Finnish literary prose from multiple Indo-European and Finno-Ugric source languages[3] (henceforth KKAother). The sub-corpus SKA contains 1,212,770 word-tokens and the sub-corpus KKAen 1,410,281 word-tokens. The KKAen sub-corpus, which contains translations by multiple translators, is needed to control for the idiolect of a single translator.

I picked the first two books from the KKAru corpus and combined them with the entirety of the KKAother corpus. The result is my multi-source-language corpus (henceforth KKAmulti) that I utilize for controlling for source language influence. I did not include the entirety of KKAru in order not to give disproportionate weight on Russian as a source language. This combined corpus contains 1,148,215 word-tokens.

The process of using a third corpus to control for source language influence mirrors Jantunen's (2004: 106–108) material choices in his method called the Three-Phase Comparative Analysis. Unlike Jantunen, I did not include English as one of the source languages in my KKAmulti corpus, because the corpus is not compared to native Finnish without the CEALfi corpus and, thus, does not need to represent source language independent translated Finnish in itself (see my mathematical test for interesting n-grams in Chapter 4.5). If English were included in the KKAmulti corpus, it would make it easier, unnecessarily, for an n-gram to pass the control as a significant part of the control corpus would be of the same proposed language variant (Finnish literary prose translated from English) than the corpus being controlled.

4.3.3 Summary of the sub-divisions of the corpora

The corpora I use are summarized in Table 2 below.

[3] Indo-European: German, French, Dutch, Norwegian, Swedish
Finno-Ugric: Estonian, Hungarian

abbreviation	use/contents	size (word-tokens)
CEALfi	main corpus of Finnish translated from English	502,062
SKA	main corpus of native Finnish	1,212,770
KKAmulti	control corpus of Finnish translated from multiple source languages (includes KKAother as well as 2 books from KKAru)	1,148,215
KKAen	control corpus of Finnish translated from English by multiple translators	1,410,281
CEALen	control corpus of source texts of CEALfi	657,986

Table 2: Summary of the corpora used

As seen in the table above, all the corpora are of the same or adjacent orders of magnitude, and the largest corpus (KKAen) is roughly 2.8 times as big as the smallest corpus (CEALfi). As my difference assessment method is not sensitive to corpus size, the corpora are probably similar enough in size.

4.3.4 Preparing the corpora

For preparing the data, I followed the same basic principle laid out in Chapter 4.2. My workflow is in Figure 3 below.

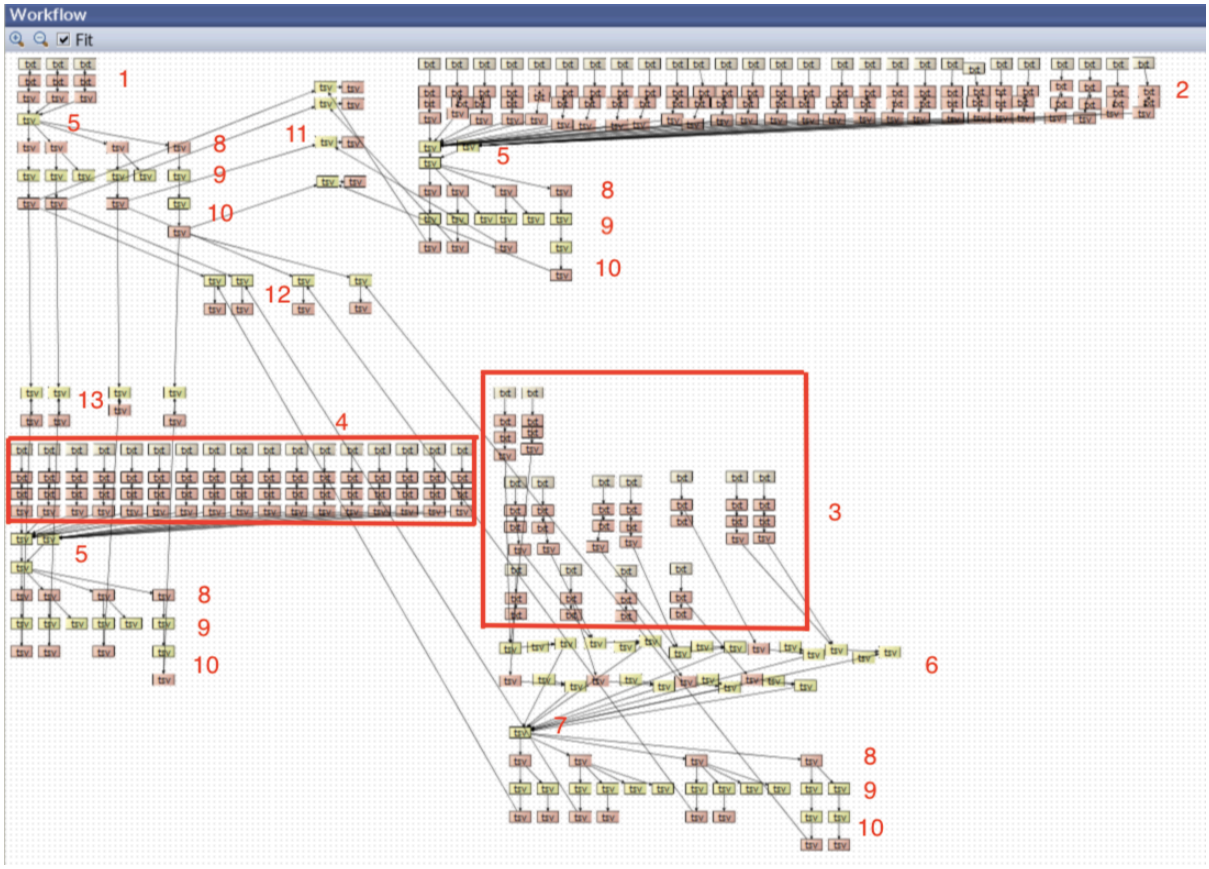

Figure 3: Workflow of corpus preparation for Finnish corpora

At number 1, there is the CEALfi corpus. The books have been parsed with Universal Dependencies 2 parser (see Universal Dependencies, a) for Finnish. I chose the Universal Dependencies framework, because it is available in Mylly and it is cross-linguistic, i.e. the same tagsets are used for all languages (Universal Dependencies, b).

At number 2, there is the SKA corpus. The files have been parsed with the Universal Dependencies 2 parser.

At number 3, there is the corpus of control translations that contains the entirety of the KKAother corpus as well as two books from the KKAru corpus. I shall refer to this corpus as KKAmulti. Every group of files represents one source language and every beige txt file one book. The books have been parsed with the Universal Dependencies 2 parser, which has produced the red txt and tsv files.

At number 4, there is the KKAen corpus that functions as a separate control corpus for ruling out features of Kersti Juva's idiolect. The books have been parsed with the Universal Dependencies 2 parser, similarly to the previous steps.

The step at number 5 (yellow files) is the summation of the CEALfi books, the summation of the SKA books, and the summation of the KKAen books. At SKA and KKAen, the summation is done in two phases, because the summation tool could only process up to 16 files at the same time, and the corpora contain more books than that.

The step at number 6 (mostly yellow files) is the summation of the parsed books of the KKAmulti corpus that have the same source language. I created a tiny relation with the language code for each sum file and joined the corresponding tiny relations and sum files. The result was that I had a sum for each source language where the source language information was included as a parameter called *origin*.

The single (yellow) file at number 7 is the sum of the sums of the translations per source language. The parameter *origin* remains in the data.

The step at number 8 (red files) is where the calculation of the n-grams occurs. As seen in the figure, the exact same procedure is carried out for each corpus. The leftmost file in each row of n-grams is for unigrams, the next for bigrams, then trigrams, and, lastly, the rightmost is for 4-grams, although I am only interested in 3+1-grams, where the fourth member is the final word-token in a sentence, nearly always a sentence-terminal punctuation mark. This decision was made because, after testing, I discovered that the distribution of n-grams where $n > 3$ is quite scattered and not nearly as informative as the distribution of, say, trigrams.

The step at number 9 (yellow files) is the calculation of the absolute frequencies of the different n-grams. In the CEALfi, SKA, and KKAen corpora, there are two files for each n-gram file (of step 8) for 2–3-grams. That is because I wanted to count not only the plain frequencies but also such frequencies where the parameter *end* differentiates the frequencies. The parameter *end* has four possible values: 0, 1, 2, and 3. The value 0 means that the gram does not touch any sentence border. The value 1 means that the first member of the gram is at the beginning of a sentence.

The value 2 means that the last member of the gram is at the end of a sentence. Lastly, the value 3 means that the gram is a whole sentence in itself.

For the 4-grams (or 3+1-grams), I only calculated the frequencies differentiated by *end*, because, as mentioned above, I am only interested in the 3+1-grams where the fourth member is a sentence-terminal punctuation mark, not in plain 4-grams. The second yellow files under the calculations are there, because the 4-grams that did not have the *end* value 2 had to be eliminated before further calculations. Unfortunately, the cases where the 4-gram is a sentence in itself, i.e. it has an *end* value of 3, could not be easily included due to the way in which Mylly does relation algebra.

In the KKAmulti corpus, there are four files for each n-gram file (of step 8) for 2–3-grams and two files for the unigram and 4-gram or 3+1-gram files. That is because I wanted the plain frequencies as well as frequencies differentiated by *end*, *origin*, and both. For the 4-gram or 3+1-gram file, I only calculated the frequencies differentiated by *end* and by *end* and *origin*. For the unigram file, I only calculated the plain frequencies and frequencies differentiated by *origin*.

The step at number 10 (red files) is the extension of the absolute frequency files with relative frequencies. The absolute frequencies of 2–3-grams differentiated by *end* were still unextended when the picture was taken, because the unwanted *end* values have to be filtered out before extending with relative frequencies in order for the sum of the relative frequencies of the grams of the same *end* value to total 100 percent. I was not yet sure which *end* values were interesting, so I left my options open.

In the KKAmulti corpus, the files that contained absolute frequencies differentiated by *origin* were extended with proportions in such a way that the relative frequencies were grouped by source language, i.e. all gram frequencies with the same *origin* value in such a file total 100 percent.

At number 11, there are the joined relations (yellow files) between files of corresponding n value and differentiation (of step 10) between the CEALfi and SKA corpora. The files contain every gram that occurs at least once in both corpora on a separate row. The files contain the parameters *cMcount*(*1*) (absolute frequency with the number differentiating the different corpora), *wMcount*(*1*) (relative frequency), and *end* (if the source files contain that parameter).

The joined relations are then extended with difference (red files). The tool produces four new parameters: the difference between the *wMcount*s, the absolute value of the difference between the *wMcount*s, the (base 10) logarithm of the ratio of the *wMcount*s, and the absolute value of the (base 10) logarithm of the ratio of the *wMcount*s.

The same procedure is also carried out between the CEALfi and KKAmulti corpora (number 12) and between the CEALfi and KKAen corpora (number 13).

4.4 Finding meaningful differences

In this section, I present two methods of finding meaningful differences between two sets of relative frequencies. These methods do not yet take the tertiary control corpora into account, for these are only the starting points. The first method is the one that Borin and Prütz (2001) use and the second method the one that I use.

4.4.1 Rank number difference

In their article, Borin and Prütz list all the n-grams of the same n value and order the list by relative frequency. They give each n-gram a rank number: the most frequent gram receives rank number 1, the second most frequent number 2 and so on. Then, they subtract the rank number of each gram from the corresponding rank number of the same gram in a different corpus. They decided to only look at grams whose rank number difference is at least 30 between their main corpora. They have multiple additional means of narrowing down what grams were interesting involving their control corpora, but I will not delve into those here. (Borin and Prütz 2001: 36–37.)

Mauranen uses a similar method when comparing word-form frequencies between corpora. She ranks the word-forms of her native Finnish corpus in descending order, excludes proper names, and divides the list into three frequency bands. Then, she does the same to her three translational corpora and calculates the sum of the rank number differences in each of the frequency bands in order to compare the relative distances between her corpora (Mauranen 2004: 75–76.)

4.4.2 Calculating logarithms of ratios

My method is to calculate the logarithm of the ratio between the relative frequencies of the same gram in two different corpora. The greater the absolute value of the logarithm, the more significant the difference. The sign of the logarithm signifies the direction of the difference.

Using a logarithm of ratio is better than a simple subtraction of the relative frequencies, because it takes the position of the gram in the distribution into account, e.g. relative frequencies 0.2% and 0.1% signify a more important phenomenon than relative frequencies 15.1% and 15.0%[4].

4.5 Bringing in the control corpus

The aim of this study is to find universal properties of translated Finnish. The differences of relative frequencies of n-grams between the CEALfi and SKA corpora may be unique to the English–Finnish language pair. This is the reason the KKAmulti corpus containing translations from other source languages is needed.

We are looking for phenomena that would, at the same time, differ greatly between the CEALfi and SKA corpora and not differ much between the CEALfi and KKAmulti corpora, because they would be most likely to indicate real differences between native and translated Finnish irrespective of source language, i.e. translation universals. I shall explain the process with the help of Figure 4 below.

[4] I thank Jussi Piitulainen, one of my instructors, for giving me this idea and helping out.

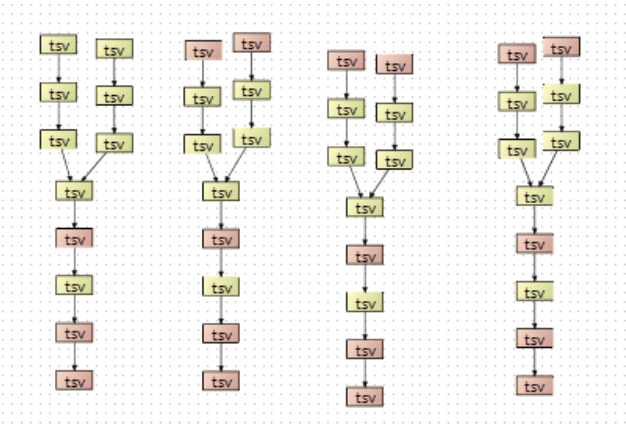

Figure 4: The process of controlling with KKAmulti

There are four pillars of files in the figure. Each pillar represents one n value, the leftmost pillar being for unigrams and the rightmost for 3+1-grams.

After calculating the differences in the n-gram relative frequencies between these corpora (Chapter 4.3.3, steps 11–12), I compiled files that contained all n-grams of the same n value and the absolute values of the logarithms of the ratios between the relative frequencies of said n-grams in a) the CEALfi and SKA corpora and b) the CEALfi and KKAmulti corpora. These files are the topmost layer of each pillar in Figure 4.

The second layer from the top is a step where I discarded all unneeded parameters from the files. The third layer is where I renamed the relative frequency parameters of the two files so that they would not be identically named. The fourth layer is the joining of the two branches into one.

I extended these files with proportions so that I received logarithms of ratios where the numerator was the value of the logarithm between the CEALfi and SKA corpora and the denominator the value of the logarithm between the CEALfi and KKAmulti corpora. In other words, the expression is as follows,

$$lg\left(\frac{\left|lg\left(\frac{x}{y}\right)\right|}{\left|lg\left(\frac{x}{z}\right)\right|}\right)$$

Figure 5: A mathematical expression for comparing differences

where x is the relative frequency of the n-gram in the CEALfi corpus, y the relative frequency of the n-gram in the SKA corpus, and z the relative frequency of the n-gram in the KKAmulti corpus. I did not calculate absolute values of this logarithm, because, here, the direction of the difference matters.

Thus, if the relative frequencies of an n-gram differ greatly between the CEALfi and SKA corpora, the numerator receives a large value, and if the relative frequencies do not differ much between the CEALfi and KKAmulti corpora, the denominator receives a small value. If both happen at the same time, the total logarithm value is large.

This combining step is the fifth layer from the top in the pillars in Figure 4. The sixth layer is just for dropping the operations the *extend with difference* tool produces that are other than logarithms of ratios.

I sorted the files from the largest logarithm values to the smallest (seventh layer in the pillars) and gave the n-grams a final interest rank according to their position in their respective lists (the bottom layer of the pillars). The twenty most interesting n-grams of each n value are included as tables.

5 Results

In this chapter, I shall go through the results I obtained and see whether they have any value as evidence for or against the proposed universals. I shall also delve into possible sources of errors. I am going to focus on the 20 most interesting n-grams unless I have a specific reason to pick one out from outside the top 20.

5.1 Unigrams

The unigrams are the same thing as individual words, so the unigram frequency table (Table 3) shows the relative frequencies of every part of speech.

kMid	unigram	SKA	CEALfi	KKAmulti	KKAen
1	VERB	0.16433289	0.15012887	0.15156395	0.15689781
2	CCONJ	0.04188098	0.04375157	0.04440806	0.04098829
3	PRON	0.09803837	0.1231561	0.11036957	0.11789991
4	NOUN	0.2342365	0.18405695	0.21310817	0.20101597
5	INTJ	0.00156831	0.00152571	0.00149275	0.00148836
6	SCONJ	0.02362773	0.03254578	0.02442835	0.02657343
7	AUX	0.06249825	0.07217834	0.06290198	0.06968966
8	PUNCT	0.17588496	0.19645183	0.17637638	0.18725346
9	ADJ	0.05784774	0.05406304	0.05797259	0.05431187
10	ADP	0.01451058	0.01257813	0.01606842	0.01443613
11	ADV	0.08309325	0.08884361	0.10027216	0.08621757
12	X	0.00061265	0.00065331	0.00073854	0.00049777
13	NUM	0.00524584	0.00425445	0.0069386	0.00593073
14	PROPN	0.03645209	0.03564699	0.03327774	0.0366622
15	SYM	0.00016986	0.00016532	8.2737E-05	0.00013685

Table 3: Unigrams sorted by interest[5]

In the table above, there are only 15 unigrams, because the Universal Dependencies 2 tagset consists of 17 different part-of-speech categories (Universal Dependencies, c), of which 15 occur in Finnish and thus at least once in all the three corpora featured in the expression in Figure 5. The two Universal Dependencies categories that do not appear in Finnish are DET and PART, or determiners and particles. Note that the notion of particle is different from traditional Finnish grammar. Most of the words that traditional Finnish grammar would count as particles are either adverbs or some kind of conjunctions in Universal Dependencies.

[5] Legend:

- kMidinterest is the ordinal number of interest according to the expression in Figure 5
- unigram is the actual part-of-speech in question
- SKA is the relative frequency of the unigram in the SKA corpus
- CEALfi is the relative frequency of the unigram in the CEALfi corpus
- KKAmulti is the relative frequency of the unigram in the KKAmulti corpus
- KKAen is the relative frequency of the unigram in the KKAen corpus

According to the table, verbs are the most consistently different part of speech between translated and native Finnish. Verbs make up approximately 16.43% of native Finnish, while the figure is 15.01% in the CEALfi corpus, 15.16% in the KKAmulti corpus, and 15.69% in the KKAen corpus. Thus, verbs seem to be markedly more frequent in native Finnish than translated Finnish. When we turn to the KKAmulti figures differentiated by source language, i.e. *origin* (see Chapter 4.3.4), we see that there are three exceptions: the Finnish translated from French (16.48%), from Norwegian (16.87%), and from Swedish (17.07%) actually contain more verbs than native Finnish. It must be noted, though, that these figures are calculated from one or two books each and are not nearly as reliable as the figures of the larger corpora. The percentages differ from those of Heikkinen et al. (2001) because Heikkinen et al. did not count punctuation marks as words.

The second part-of-speech category in the ranks is CCONJ, that is co-ordinating conjunction. At closer inspection, the finding has to be dismissed, because even though the category is more frequent in the translation corpora CEALfi and KKAmulti than in the native Finnish SKA, the KKAen figure falls below all the other figures. In other words, the unigram fails the control where the effect of single translators is mitigated in Finnish translated from English (The KKAmulti corpus does include multiple translators).

The third part-of-speech category is that of pronouns. Pronouns seem to be more frequent in translated Finnish than native Finnish (9.80% in SKA, 12.31% in CEALfi, 11.03% in KKAmulti, 11.79% in KKAen). Also, all the different figures of KKAmulti differentiated by *origin* are greater than the figure of native Finnish. This finding is in line with Mauranen and Tiittula's (2005: 42) finding that the first person singular pronoun MINÄ is more frequent in translations than in original texts. In addition to that, Auvinen (2005: 77) states, in passing, that the second person singular pronoun SINÄ is more frequent in translated Finnish than native Finnish. To confirm that these pronouns increase in frequency when comparing translations to original texts, I calculated the relative frequency of every lemma in the CEALfi and SKA corpora and looked at what pronouns' relative frequency increases the most moving from native texts (SKA) to translations (CEALfi). The top 5 is HÄN, EI, MINÄ, JOKA, SINÄ. Thus, Mauranen and Tiittula's (2005: 42) and Auvinen's (2005: 77)

findings gain support from my data. It has to be noted, though, that the lemma EI is not a pronoun but an auxiliary verb, but due to the way I compiled the list, it is enough that a single occurrence of EI is (falsely) tagged as a pronoun for the whole increase of the relative frequency of the lemma making it to the list, regardless of the actual taggings of individual occurrences.

The phenomenon of more frequent personal pronouns in translations could be said to support at least three universals: *interference*, *under-representation of unique items*, and (*T*-)*explicitation*. The first two, *interference* and *under-representation of unique items* are two sides of the same coin. In Finnish, a genitive personal pronoun modifying a noun and denoting possession can be left out, because the relation is visible in the possessive suffix of the head noun. For example, both *minun autoni* and *autoni* ('my car') are perfectly fine constructions. As the pronoun is visible in the source text, the translator might often let it stay there in the translation, as the unique possibility of leaving the pronoun does not suggest itself as an equivalent (cf. Tirkkonen-Condit 2004: 177–178). The third universal, (*T*-)*explicitation*, can be said to receive support, because the translations have more cases of double explicitation of the relation of possession between the pronoun and noun.

The part-of-speech category of nouns also seems to be consistently different in frequency between native and translated Finnish, namely nouns being more frequent in native Finnish (23.42% in SKA vs. 18.41% in CEALfi, 21.31 in KKAmulti, and 20.10% in KKAen). Kersti Juva's translations in particular seem to be noun-sparse. When looking at the KKAmulti figures differentiated by *origin*, we see that there are two exceptions to the general tendency: the Finnish translated from French (24.46%) and from Dutch (24.20%) actually contain more nouns than native Finnish.

These figures are in the same ballpark as Hudson's (1994). He claims that the share of (common) nouns is somewhat similar (around, perhaps a little over, 20%) in all written language regardless of the genre or actual language in question. Hudson also gives more specific figures of 19% and 17% in "imaginative" (i.e. fictive) English in the Brown and LOB corpora, respectively (Hudson 1994: 332). However, the percentage of nouns in the CEALen corpus is only 12.69%, which raises the question of whether punctuation is included in Hudson's part-of-speech categorization (no information on this is given in the research note). If the categorization differs in

regard to punctuation, the figures between this thesis and Hudson's paper are not comparable. It has to be noted, though, that the relative frequency of (common) nouns without punctuation marks as word-tokens is still only 15.2% in the CEALen corpus.

What Hudson (1994: 332, 336–337) also brings up and what is more resistant to differences in tagsets is the observation that the frequency of (common) nouns is inversely correlated with the frequency of pronouns in many languages. This tendency can be seen in my data, as pronouns are more frequent and common nouns less frequent in translations than in original texts, as shown above.

In addition to the frequency of pronouns, Hudson (1994: 336–337) writes that the frequency of verbs seems to be connected to that of (common) nouns. More nouns should yield less verbs. This tendency gets no support from my data, as both verbs and nouns are more frequent in native Finnish than translated Finnish.

The next part-of-speech category in the list is interjections. The frequency differs noticeably between the different translation corpora, but all the translation figures fall below the native figure (1.57‰ in SKA vs. 1.53‰ in CEALfi, 1.49‰ in KKAmulti, and 1.49‰ in KKAen). There are two exceptions to this when we look at the KKAmulti figures differentiated by *origin*: the Finnish translated from Estonian (3.15‰) and from Norwegian (2.05‰) contain more interjections than native Finnish. As interjections are colloquial in nature, this finding could be said to be in support of the *conventionality* universal, as interjections, which are as a category quite unconventional in written language, are less frequent in translations, despite the two outlier source languages. The finding is also in line with Puurtinen's (2005) findings of less colloquial language in translations than in native texts.

The sixth part-of-speech category is that of subordinating conjunctions. Their frequency differs quite a lot between the different translation corpora, but all the translation figures seem to be greater than the native figure (2.36% in SKA vs. 3.25% in CEALfi, 2.44% in KKAmulti, and 2.66% in KKAen) until we look at the figures of KKAmulti differentiated by *origin*. Then we see that the figures of Finnish translated from Estonian (2.27%), Spanish (2.29%), French (1.88%), and Swedish (2.36%) fall below the native figure. Despite the figures differentiated by *origin*, the

differences in the frequency of subordinating conjunctions raises a new hypothesis of more infinitival structures in native Finnish than in translated Finnish (at least from English), because most infinitival structures are interchangeable with a corresponding subordinate clause. Such phenomenon has already been partially supported by Eskola (2005).

The seventh part-of-speech category is auxiliary verbs. In Finnish, these include

- "*täytyä* 'must'
- *pitää* 'have to'
- *tarvita* 'need'
- *joutua* 'have to'
- *voida* 'be able to, can'
- *saattaa* 'may'
- *taitaa* 'be+probably, may'
- *mahtaa* 'be+probably, may'
- *olla* 'be'
- *aikoa* 'be going to'" (Universal Dependencies, d).

The frequency of auxiliary verbs differs quite a lot between the translation corpora, but all the translation figures seem to be greater than the native figure (6.25% in SKA vs. 7.22% in CEALfi, 6.29% in KKAmulti, and 6.97% in KKAen) until we look at the figures differentiated by *origin*. Then we see that the figures of Finnish translated from German (6.22%), Estonian (5.73%), French (5.51%), and Russian (6.00%) fall below that of native Finnish.

The last part-of-speech category that passes the main test (see Figure 4) is punctuation. The translations have more punctuation in them than the original texts (17.59% in SKA vs. 19.65% in CEALfi, 17.64% in KKAmulti, and 18.73% in KKAen). Again, some individual source languages are outliers, namely Estonian (17.54%), French (16.47%), and Dutch (15.91%), but the same problem of small corpus size remains. Despite the outliers and although punctuation is used not only at clause borders, to me, the finding as a whole suggests that the translations have, on average, shorter clauses than the native texts. This could be said to support the *simplification* universal if one regards shorter clauses as simpler clauses.

After punctuation, the remaining part-of-speech categories fail the main test, i.e. the difference between the CEALfi and KKAmulti corpora is more notable than the difference between the CEALfi and SKA corpora.

To conclude, the most reliable finding from these unigrams is the fact that translated Finnish contains more pronouns than native Finnish. In addition to that, the frequencies of verbs, interjections, subordinating conjunctions, auxiliary verbs, and punctuation have quite reliable tendencies to being either over- or under-represented in translated Finnish.

5.2 Bigrams

The bigrams are strings of two consecutive word-tokens. The twenty most interesting bigrams according to the test in Figure 4 are in Table 4 below.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
1	VERB PUNCT	0.03739125	0.03116023	0.03117582	0.03199152
2	AUX INTJ	4.6141E-05	6.3553E-06	6.5967E-06	1.621E-05
3	VERB CCONJ	0.00521663	0.00365644	0.00370073	0.00416818
4	ADJ NUM	0.00010495	0.00014194	0.00014041	0.00016904
5	INTJ AUX	3.5284E-05	2.1184E-05	2.0732E-05	2.0841E-05
6	PROPN INTJ	1.0857E-05	2.1184E-06	1.8848E-06	1.8525E-05
7	ADP PRON	0.00095267	0.00126471	0.00129295	0.0012875
8	AUX ADV	0.01191971	0.01517655	0.01488772	0.01494136
9	PUNCT X	0.00013299	0.00021396	0.00020544	0.00015978
10	PRON ADV	0.00759966	0.00979146	0.01000996	0.00817039
11	PRON ADP	0.00227085	0.00322004	0.00311645	0.00311378
12	SCONJ ADV	0.00166107	0.00220107	0.00214015	0.00155766
13	X PUNCT	0.00018818	0.00026904	0.00025821	0.00023002
14	INTJ SCONJ	3.7998E-05	3.1777E-05	3.1099E-05	1.3894E-05
15	SCONJ INTJ	1.8094E-05	2.1184E-06	2.8271E-06	7.7188E-07
16	CCONJ SYM	9.0472E-07	2.1184E-06	1.8848E-06	2.3157E-06
17	PROPN SCONJ	7.1473E-05	5.508E-05	5.7485E-05	8.7223E-05
18	X CCONJ	7.2378E-06	1.2711E-05	1.1309E-05	1.3894E-05
19	PRON ADJ	0.00521029	0.00653964	0.00620275	0.00543715
20	PRON NOUN	0.02503908	0.0282283	0.02738179	0.02648875

Table 4: The 20 most interesting bigrams

In the following paragraphs, we study the results bigram by bigram and make tentative conjectures about the nature of translated Finnish.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
1	VERB PUNCT	0.03739125	0.03116023	0.03117582	0.03199152

The first bigram is VERB PUNCT. As punctuation marks are often on clause borders, to me, the VERB PUNCT frequencies suggest that native Finnish has clause-ending verbs more often than translated Finnish. To check this hypothesis, I calculated the relative frequencies of all bigrams differentiated by *end* with *end* value 2 (i.e. all sentence-terminal bigrams). Of those sentence-terminal bigrams, VERB PUNCT makes up 21.34% in SKA, 14.22% in CEALfi, 16.24% in KKAmulti, and 17.45% in KKAen, so my guess was correct. However, it has to be mentioned that not all verbs are finite, so a portion of these verbs might be participles or infinitivals functioning not unlike a noun or an adjective.

The verb-final construction is marked in nature when the verb follows its qualifiers and there is something before the theme position in the clause. What is common to most of those marked constructions is that they convey reactions and affections, and, thus, are somewhat colloquial. (VISK § 1390.) If the marked, colloquial constructions are more frequent in native texts than in translations, it would support the *conventionality* universal. Unfortunately, I am not able to do theme–rheme analysis automatically, so the hypothesis cannot be properly tested. See, however Chapter 5.6.

When we look at the bigrams differentiated by *origin*, we see that there are two source languages that are outliers, namely Norwegian and Swedish. VERB PUNCT makes up 4.78% and 4.78% of all bigrams in Finnish translated from those source languages, respectively, as well as 24.84% and 24.88% of sentence-terminal bigrams (*end* value 2), respectively. All the other source-language differentiated figures fall below the native Finnish figure in both categories. Thus, it seems that Scandinavian languages have something in them that triggers verb-final clauses in Finnish translations. It would be interesting to see whether this holds true with Finnish literary prose translated from, e.g., Danish and Icelandic as well, but the CTF does not include those source languages.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
2	AUX INTJ	4.6141E-05	6.3553E-06	6.5967E-06	1.621E-05

The second most interesting bigram is AUX INTJ. In addition to the translation figures above, all the KKAmulti figures differentiated by *origin* fall below the native figure. The absolute majority of the individual cases are either *ei* ('no/not') or some form of *olla* ('be') followed by some expletive. A very typical example is *ei vittu* (roughly the same in function as *oh fuck*). The finding supports the notion of translations being more conservative and less colloquial in nature (and thus the *conventionality* universal gains support). The findings are in line with the earlier observation of less interjections *per se* in translations, as well as Puurtinen's (2005) findings of less colloquial language in translations than in native texts. It has to be noted, though, that the CEALen corpus of English source texts does not contain a single common expletive, so the difference may be due to the novels itself being different.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
3	VERB CCONJ	0.00521663	0.00365644	0.00370073	0.00416818

The next bigram in the list is VERB CCONJ. In addition to the translation figures above, all the KKAmulti figures differentiated by *origin* fall below the native figure as well. At first glance, this finding seems to be a manifestation of the same phenomenon as with the bigram VERB PUNCT, namely clauses ending in verbs in native Finnish (this time the clause being followed by a co-ordinating conjunction instead of a punctuation mark).

The absolute majority of the co-ordinating conjunctions are *ja* ('and'), so I wanted to see whether the *ja* was followed by a co-ordinated verb or another clause. I picked out all *trigrams* beginning with VERB CCONJ and found out that the third member is another verb in a little over half of the cases. I also calculated *5-grams* (see step 8 in Chapter 4.3.4.) from the corpora and picked out all that begin with VERB CCONJ to see whether the sentences go on after the conjunction. If there is a 5-gram to be found, the shortest possibility should be VERB CCONJ A B PUNCT, where A is a verb in over 50% of cases. In other words, if there is a corresponding 5-gram, the clause after the conjunction is at least two words long. The results are that there is a 5-gram in approximately 90.48% of the cases in SKA, 94.79% in CEALfi, 92.69% in

KKAmulti, and 92.74% in KKAen. Thus, it can be concluded that 1) it is not very common for a sentence to end with a co-ordinating conjunction followed by 0–1 words 2) the finding is mostly a secondary manifestation of the phenomenon described in connection with the VERB PUNCT bigram, namely native Finnish clauses ending in verbs more often that translated Finnish clauses.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
4	ADJ NUM	0.00010495	0.00014194	0.00014041	0.00016904
157	NUM ADJ	0.00046774	0.00037497	0.00057014	0.00050558

The fourth bigram is ADJ NUM, i.e. an adjective followed by a numeral. Despite the general tendency being that the bigram is more frequent in translations, we see that there are four outlier source languages in the KKAmulti figures differentiated by *origin*, namely German (0.0901‰), Norwegian (0.0629‰), Russian (0.1046‰), and Swedish (0.0176‰). It could be that the construction of the type 'the last two' that can be realized as either ADJ NUM or NUM ADJ (*viimeiset kaksi* vs. *kaksi viimeistä*) would be more often rendered as ADJ NUM in translations than in original texts, but when we look at the bigram NUM ADJ, we see that it is also more frequent in the KKAmulti and KKAen corpora than in SKA (the CEALfi being an outlier in this case, explaining the gram's relatively low placement in the ranks). Thereby, the bigrams fail to show us anything meaningful. It has to be noted that the trigram NUM ADJ CCONJ appears fourth in the trigram interest rankings (see Table 5) with the gram being more frequent in native texts, but the individual occurrences of that gram are not of the type 'the last two'.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
5	INTJ AUX	3.5284E-05	2.1184E-05	2.0732E-05	2.0841E-05

The next bigram is INTJ AUX, which is more frequent in native texts than in translations. However, there are three outlier source languages, namely Estonian (0.0393‰), Spanish (0.0362‰), and Russian (0.0374‰). Most of the actual occurrences are either *no on* (discourse particle + '[it] is'), *no ei* (discourse particle + 'no'), or *voi ei* ('oh no'). In addition to these, in the native Finnish data, there are expletives such as *vittu ollu* and *perkele on*. Again, we see that the expletives are a trait of native Finnish and native Finnish only. Actually, when we subtract all the expletive + AUX occurrences, the relative frequency of the bigram in native Finnish

becomes 0.0280‰, which is more in line with translated language. Thus, the *conventionality* universal gains support once again, and Puurtinen's (2005) findings are replicated.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
6	PROPN INTJ	1.0857E-05	2.1184E-06	1.8848E-06	1.8525E-05

The sixth bigram is PROPN INTJ, i.e. a proper noun followed by an interjection. The absolute frequencies of the bigram in the translations are too low to say anything with certainty, as there are only 1 such bigram in all CEALfi, and 2 in KKAmulti (one of which is a clear tagging error). What can be said is that the most common occurrence is [PROPN] *hei* ('Hey [PROPN]') and that the native Finnish data contains a couple of expletives.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
7	ADP PRON	0.00095267	0.00126471	0.00129295	0.0012875

The next bigram is ADP PRON, an adposition followed by a pronoun. The bigram is generally more frequent in translations than in native texts. When looking at the KKAmulti figures differentiated by *origin*, there are two exceptions to the bigram being more frequent in translations: in Finnish translated from Norwegian (0.803‰) and from Swedish (0.826‰), the bigram is less frequent than in native Finnish. However, when looking at the individual occurrences, the bigram does not reflect any syntactical structure, as there is virtually always a phrase border between the constituents. It could have been that the few Finnish adpositions that can function as both pre- and postpositions would be rendered more often as prepositions in translations, but it does not seem to be the case. The finding may simply reflect the general tendency of more pronouns in translations than in native texts.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
8	AUX ADV	0.01191971	0.01517655	0.01488772	0.01494136

The eight bigram is AUX ADV. In addition to the translation figures above, all the KKAmulti figures differentiated by *origin* are greater than the native figure. Thus, it seems that the bigram is clearly more frequent in translated Finnish than in native Finnish. The vast majority of the verbs are some forms of *olla* ('be').

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
9	PUNCT X	0.00013299	0.00021396	0.00020544	0.00015978

The ninth bigram is PUNCT X. X is a category where all the words that cannot be assigned a real part-of-speech category fall into. In most cases, such words are snippets of other languages embedded in the text. (Universal Dependencies, e.) The bigram seems to be more frequent in translated language than in native Finnish. Most of the X category words in the data are foreign words, mostly English but also French and German. There are also some semi-wild expressive words such as *tsuiikk*, *pst*, and *iih*. Semi-wild expressive words are words whose phonetic (and thereby also orthographic) form mirrors some quality (most often sound) of their referent. They are not established as a recurrent part of the language but are only used *ad hoc*. (Jääskeläinen 2015: 464, 466–467.) In the KKAmulti figures differentiated by *origin*, there are three outlier source languages whose figure falls below the native figure, namely Hungarian (0.0764‰), Dutch (0.0527‰), and Swedish (0.0351‰), and one outlier language to the other direction, namely Estonian (0.4102‰).

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
10	PRON ADV	0.00759966	0.00979146	0.01000996	0.00817039

The tenth bigram in the list is PRON ADV. The bigram is more frequent in translations than in native texts. In the KKAmulti figures differentiated by *origin*, there are two outlier source languages, namely Dutch (0.7444%) and Swedish (0.6587%). When looking at the individual occurrences, no clear syntactic pattern emerges. The finding may simply reflect translations' having more pronouns in them in general than native texts (see Chapter 5.1).

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
11	PRON ADP	0.00227085	0.00322004	0.00311645	0.00311378

The next bigram, PRON ADP, also has to do with pronouns. It is more frequent in translated texts than in native texts. When looking at the KKAmulti figures differentiated by *origin*, we find two source languages whose respective figures fall below the native figure, but only by a very small margin. The languages are French (2.229‰) and Swedish (2.231‰). Again, when looking at the individual

occurrences, no single, clear-cut syntactical pattern emerges. Again, I suspect that the finding simply reflects the general commonness of pronouns in translated Finnish.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
12	SCONJ ADV	0.00166107	0.00220107	0.00214015	0.00155766

The twelfth bigram in the list is SCONJ ADV. Even though the bigram seems to be generally more frequent in translations than in native texts, the KKAmulti figures differentiated by *origin* differ wildly from 2.864‰ (Hungarian) down to 1.071‰ (Swedish). The syntactical pattern the bigram mirrors is a subordinate clause beginning with an adverb(ial). No other pattern of distribution can be found than the frequency of the bigram differing wildly between different datasets, possibly reflecting individual writers' preferences (the KKAmulti *origin*-differentiated sub-corpora consisting of 1–2 books each).

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
13	X PUNCT	0.00018818	0.00026904	0.00025821	0.00023002

The next bigram, X PUNCT, brings us back to the class X. Generally, the bigram is more frequent in translations than in native texts. In the KKAmulti figures differentiated by *origin*, there are three languages whose respective figures fall below the native figure: Spanish (0.1085‰), Dutch (0.0702‰), and Swedish (0.0176‰). There is also one outlier to the other direction, namely Estonian (0.4931‰). The occurrences of X are of the same kind as with PUNCT X: mostly English, German, and French words as well as semi-wild expressive words. There are also quite a few tagging errors. However, as the general direction of the difference is the same and the outlier source languages are almost exactly the same as with the PUNCT X bigram earlier, these X words might deserve further study.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
14	INTJ SCONJ	3.7998E-05	3.1777E-05	3.1099E-05	1.3894E-05
15	SCONJ INTJ	1.8094E-05	2.1184E-06	2.8271E-06	7.7188E-07

The next two bigrams have again to do with interjections, first, before a subordinate clause (INTJ SCONJ) and, second, the other way around (SCONJ INTJ). Both are more frequent in native texts. In addition, all the KKAmulti figures differentiated by *origin* are in line with the general tendency, with some source languages actually

showing 0 absolute occurrences. The INTJ SCONJ occurrences in the translations are mostly either *Voi kun*, *Kas kun*, *No kun*, or *Ai että*. The native occurrences are of the same type, but, again, in addition to those, there are expletives such as *vittu kun* and *saatana kun*. The SCONJ INTJ occurrences are very rare, and almost all of the occurrences in translations are tagging errors. The native occurrences are *että hei* (4 occurrences), *että no* (2), *jos meinaan* (1), and some subordinating conjunction + an expletive (13 occurrences). In comparison, in both the bigrams, there are exactly three expletives in translation, all of which in the KKAen corpus. Again, the tendency of translations to become *conventional* and less colloquial than native texts receives support.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
16	CCONJ SYM	9.0472E-07	2.1184E-06	1.8848E-06	2.3157E-06

The sixteenth bigram is CCONJ SYM, but all but one of its few occurrences are tagging errors, so nothing more can be said about it.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
17	PROPN SCONJ	7.1473E-05	5.508E-05	5.7485E-05	8.7223E-05

The seventeenth bigram is PROPN SCONJ. As the KKAen corpus is not in line with the other translation corpora here, no straight conclusions can be made. In addition to that, the KKAmulti figures differentiated by *origin* range all the way from 0.1352‰ (German) to 0.0117‰ (Dutch). I suspect that the variation simply reflects the very varying degree of proper nouns in different texts that has little to do with translation, which was the reason why Borin and Prütz (2001: 37) discarded the n-grams that contained proper nouns altogether from their data.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
18	X CCONJ	7.2378E-06	1.2711E-05	1.1309E-05	1.3894E-05

The eighteenth bigram in the list is X CCONJ, but virtually all of its occurrences are tagging errors.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
19	PRON ADJ	0.00521029	0.00653964	0.00620275	0.00543715

The nineteenth bigram is PRON ADJ. The bigram seems to be more frequent in translated texts than in native texts. When we turn to the KKAmulti figures differentiated by *origin*, we see that there are three source languages whose figures fall below the native figure: Norwegian (3.839‰), Spanish (5.136‰), and Swedish (5.182‰). As no clear syntactical pattern can be distinguished, I suspect the finding simply mirrors the general tendency of translations containing more pronouns than native texts.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
20	PRON NOUN	0.02503908	0.0282283	0.02738179	0.02648875

The bigram that rounds up the top 20 is PRON NOUN. The bigram seems to be more frequent in translated texts than in native texts. There is some dispersion in the KKAmulti figures differentiated by *origin*, and three source languages have their figures fall below the native figure, namely Norwegian (2.20%), Russian (2.45%), and German (2.49%). As the bigram is very common and no clear syntactical pattern emerges from the individual occurrences, I strongly suspect that this finding again simply mirrors the pronoun-richness of translated Finnish.

5.3 Trigrams

The 20 most interesting trigrams are in Table 5 below.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
1	SCONJ PUNCT ADV	7.0108E-06	2.2621E-05	2.2579E-05	1.1008E-05
2	NOUN VERB SCONJ	0.0005829	0.00031669	0.00031713	0.00063844
3	PUNCT PUNCT PROPN	0.00055987	0.00106543	0.00106327	0.00249871
4	NUM ADJ CCONJ	2.0031E-05	1.131E-05	1.129E-05	1.3548E-05
5	ADJ CCONJ SCONJ	2.0031E-05	3.3931E-05	3.3869E-05	2.8789E-05
6	PROPN PRON ADP	1.302E-05	1.131E-05	1.129E-05	2.2015E-05
7	CCONJ PRON NUM	2.3036E-05	4.9765E-05	5.029E-05	3.1329E-05
8	PROPN PROPN PRON	3.5054E-05	4.2979E-05	4.3106E-05	5.5038E-05
9	PUNCT AUX ADP	6.0093E-06	1.3572E-05	1.3342E-05	7.6206E-06
10	PUNCT PUNCT PUNCT	0.00018829	0.00057909	0.00059321	0.00052328
11	PRON PUNCT VERB	0.00061295	0.0007216	0.00072458	0.00062404
12	CCONJ PRON ADJ	0.0003215	0.00043658	0.00044029	0.000337
13	VERB SCONJ PUNCT	1.1017E-05	3.3931E-05	3.2842E-05	1.4394E-05
14	AUX PROPN PROPN	0.00018529	0.00012894	0.00013034	0.00020914
15	ADV PUNCT ADJ	0.00027643	0.00036871	0.00036537	0.00026587
16	ADP ADV CCONJ	2.3036E-05	1.3572E-05	1.3342E-05	1.0161E-05
17	PRON NUM PUNCT	3.6056E-05	4.9765E-05	5.029E-05	4.149E-05
18	PROPN SCONJ VERB	1.7026E-05	9.0482E-06	9.2369E-06	9.3141E-06
19	AUX PROPN PRON	3.4053E-05	1.8096E-05	1.8474E-05	2.7095E-05
20	PUNCT PUNCT ADV	0.00023236	0.00069445	0.00072253	0.0002972

Table 5: The 20 most interesting trigrams

From now on, I will only write about grams that yield results.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
2	NOUN VERB SCONJ	0.0005829	0.00031669	0.00031713	0.00063844

The first trigram yielding any results is NOUN VERB SCONJ, which seems to be more frequent in native texts than in translations. However, the KKAen corpus and the KKAmulti data translated from Spanish (0.6423‰) form two exceptions to this general tendency. Over half of the individual occurrences reflect the syntactic pattern subject + predicate + subordinating conjunction, e.g. "*Asia edellyttää että*" ('The matter requires that') or "*Äiti nousi kun*" ('Mom stood up when'). The two most common subordinating conjunctions in the third position are *kuin* ('like') and *että* ('that'), the former beginning a simile and the latter separating a reporting clause

from the report itself. This phenomenon might be due to other reasons than the fact that translations are translations. It might just be that the different books selected for the corpora have different amounts of similes or indirect quotes.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
3	PUNCT PUNCT PROPN	0.00055987	0.00106543	0.00106327	0.00249871

The next trigram is PUNCT PUNCT PROPN. The occurrences are most often either the end of a direct quote and the beginning of a reporting clause (quotation mark + comma + name of the person who was quoted, e.g. "*", Olli*") or the beginning of a direct quote (colon + quotation mark or quotation dash + first word of quote, e.g. "*: - Antonio*"). In the latter case, many of the first words in the quotes are not actually proper nouns (e.g. "*: - Heippa*" [':—Bye']). The tagger is being misled by the capitalization of the first letter of the word. The gram seems to be more frequent in translations than in native texts, although there are four outlier source languages in the KKAmulti figures differentiated by *origin*, namely Norwegian (0.0702‰), Hungarian (0.3347‰), Spanish (0.3613‰), and French (0.3971‰). I find it difficult to believe that translations would systematically have more direct quotes. I think the finding may reflect a difference in orthographic conventions: in native Finnish, the quotation dash seems to be used much more often than in translations. The ends of the quotes marked with quotation dashes do not contain a second punctuation mark in addition to the comma, so they do not raise the frequency.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
5	ADJ CCONJ SCONJ	2.0031E-05	3.3931E-05	3.3869E-05	2.8789E-05

The fifth trigram is ADJ CCONJ SCONJ. It seems to be more frequent in translations than in native texts (with the exception of Estonian as a source language [0.0047‰]). The CCONJ SCONJ part of the gram corresponds either with a structure where a subordinate clause is wedged in the beginning of a co-ordinated main clause or a structure where two subordinate clauses are co-ordinated with one another. A typical example of the gram is "*porvarillisempi ja jos*" ('more bourgeois and if'). However, the ADJ in the beginning does not seem to signify anything in particular. Thereby, I turned back to the bigram list and sought out the bigram CCONJ SCONJ.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
58	CCONJ SCONJ	0.00097167	0.00217141	0.00135797	0.00118253

With the ADJ removed from the beginning, we still see the same tendency. The bigram is more frequent in translated language. In the KKAmulti figures differentiated by *origin*, there are three source languages whose figures fall below the native figure, namely Norwegian (0.7553‰), Dutch (0.8368‰), and Swedish (0.8958‰). As the position between the CCONJ and SCONJ is a typical place for a comma, the finding might reflect punctuation differences in the two systems, and thus, *interference*, as the translations might retain some source punctuation characteristics. To control this, I sought out the trigram CCONJ PUNCT SCONJ.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
1555	CCONJ PUNCT SCONJ	3.0046E-6	4.5241E-6	1.8474E-5	1.5241E-5

Even though the results vary, the tendency with the punctuation mark in between the two constituents remains similar, as the gram is still more frequent in translations than in native texts. If the finding were a manifestation of punctuation *interference*, the opposite would be expected.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
7	CCONJ PRON NUM	2.3036E-05	4.9765E-05	5.029E-05	3.1329E-05

The trigram CCONJ PRON NUM seems to be more frequent in translated language than in native language. The only exception is Finnish translated from Swedish (0.0195‰). The PRON NUM part of the gram seems to correspond to a structure where the pronoun functions as a qualifier for the noun phrase beginning with the numeral. The pronoun often signifies either possession (genitive form of a personal pronoun), e.g. "*ja hänen kolme*" ('and his/her three') or definiteness, e.g. "*ja ne pari*" ('and those couple of'), in which case the pronoun is not very much unlike a definite determiner. However, as was the case above, the first member of the trigram does not seem to be syntactically connected to the phenomenon. Thus, I again turned to the bigram list and sought out the bigram PRON NUM.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
112	PRON NUM	0.00040079	0.00050207	0.00063139	0.00048937

The same tendency can be seen here. The bigram is more frequent in translated language than in native language. There are three exceptions to that in the KKAmulti figures differentiated by *origin*: Finnish translated from Norwegian (0.2675‰), from

Swedish (0.3162‰), and from Russian (0.3959‰). Why, then, does the trigram where the first member does not belong to the construction rise so much higher in the ranks than the bigram? I think this phenomenon is due to the first member of the gram disambiguating the construction by giving it context. When PRON NUM is by itself, the two words do not necessarily belong to the same phrase, but, after a conjunction, they most often do, helping the possible underlying phenomenon to be noticed. Here, despite the outliers, we might conclude that the structure where there is a pronominal qualifier to a noun phrase beginning with a numeral is more frequent in translated language than in native language.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
10	PUNCT PUNCT PUNCT	0.00018829	0.00057909	0.00059321	0.00052328

The trigram PUNCT PUNCT PUNCT seems to be more frequent in translations than in native texts. There are, however, four source language exceptions, namely Norwegian (0.0527‰), Dutch (0.0893‰), Spanish (0.1606‰), and Swedish (0.1756‰). The vast majority of the individual occurrences have at least one quotation mark in them, so the finding may be a reflection of the same phenomenon already observed in connection with the PUNCT PUNCT PROPN trigram, i.e. translations perhaps having more direct quotes than native texts. It has to be noted that many of the occurrences, e.g. *!″″*, *.″″*, and *?″″*, seem to cross sentence borders. The sentence border is in between the quotation marks. There probably is a line break there, but it does not show up in the processed files. This crossing of sentence borders should not be happening as the Mylly n-gram tool respects sentence borders and does not calculate grams that would cross said borders. It seems that the computer is confused by the amount of punctuation and gets the sentence border placement wrong.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
12	CCONJ PRON ADJ	0.0003215	0.00043658	0.00044029	0.000337

The trigram CCONJ PRON ADJ seems to be more frequent in translated language than in native language, albeit the difference between SKA and KKAen is very slim. In addition, there are four source languages in KKAmulti, whose translations have a lower frequency than native Finnish: Hungarian (0.2929‰), Norwegian (0.2808‰), Swedish (0.1951‰), and Spanish (0.1606‰). There is, however, a tendency to be

seen: there seems to be more constructions in translations where the pronoun is used to mark definiteness, e.g. "*ja tämän seuralaisen*" ('and this companion['s]'), "*ja ne mieluisat*" ('and those pleasant'), "*ja tämä lupaava*" ('and this promising'), and "*ja nämä aamuöiset*" ('and these after midnight', literally 'and these morning-nightly'). The same definiteness-marking tendency of pronouns in translations was already partially detectable from the previous gram CCONJ PRON NUM. This marking of definiteness may be a manifestation of *interference*, as Finnish does not usually mark definiteness in an explicit manner (although it has been proposed that the demonstrative pronoun *se* is being grammaticized into a definite article [Laury 1996]). If the source text has an explicitly definite noun phrase, the explicitness may carry over to the translation.

The conjunction in the beginning of the gram does not belong to the construction but it disambiguates the context for the test, i.e. after the conjunction, the two other members are more likely to belong to the same phrase, helping the possible underlying phenomenon to shine through. This is exactly what happened, since the definiteness-marking use of the pronoun was not detectable from the bigram PRON ADJ before, when the occurrences were looked at in order to see what impressions might arise from them. Also, in the previous gram pair CCONJ PRON NUM and PRON NUM, the former gram with more phrase-disambiguating context, had a higher kMid number.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
16	ADP ADV CCONJ	2.3036E-05	1.3572E-05	1.3342E-05	1.0161E-05

The gram ADP ADV CCONJ seems to be more frequent in native Finnish than in translated Finnish. However, there are two source languages in the KKAmulti figures differentiated by *origin* that form an exception: Dutch (0.0255‰) and Norwegian (0.0351‰). The gram itself seems to favor a structure where a postposition might be followed by two adverbials co-ordinated by a co-ordinating conjunction. To check this, I calculated the *tetragram* ADP ADV CCONJ ADV.

tetragram	SKA	CEALfi	KKAmulti	KKAen
ADP ADV CCONJ ADV	1.0077E-05	2.4247E-06	2.2507E-06	1.8732E-06

The same tendency remains here. The gram is more frequent in native Finnish than in translated Finnish. A typical example of the gram is "*kanssa ylös ja alas*" ('with [someone] up and down'). In addition, all the KKAmulti figures differentiated by *origin* fall below the native figure. In fact, only two source languages, German and Dutch, have any occurrences at all. To control whether the adposition in the beginning has to do with the phenomenon, I also sought out the trigram ADV CCONJ ADV.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
558	ADV CCONJ ADV	0.00063899	0.00088673	0.0007174	0.00057408

Here, the tendency disappears. The native Finnish figure is in the middle of the translation figures. Thus, either the difference in frequency is somehow tied to the adposition or the adposition disambiguates the string ADV CCONJ ADV well as belonging to the same phrase. To see, whether this is true, I sought out the sentence-terminal (*end* value 2) 3+1-gram ADV CCONJ ADV PUNCT. As the sentence border should disambiguate the construction similarly, the tendency should reappear if the phenomenon is not tied to the adposition *per se*.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
585	ADV CCONJ ADV PUNCT	0.00133989	0.00166939	0.00127186	0.00117732

The tendency does not unambiguously reappear. While the native figure is higher than the KKAmulti and KKAen figures, the CEALfi figure is even higher. As the result is very ambiguous, I do not feel comfortable drawing conclusions to one way or another.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
17	PRON NUM PUNCT	3.6056E-05	4.9765E-05	5.029E-05	4.149E-05

The next trigram is PRON NUM PUNCT. It seems to be more frequent in translations than in native texts. The gram reflects the same phenomenon already observed in connection with CCONJ PRON NUM: the bigram part PRON NUM is over-represented in translations. Not unlike in the former gram's occurrences, the pronoun acts as a qualifier for the noun phrase the numeral starts or forms by itself, signifying either possession, e.g. "*toisen kuusikymmentä.*" ('another's sixty.') or, more often, definiteness, e.g. "*se yksi,*" ('that one,'). The outlier source languages in

the KKAmulti figures differentiated by *origin* are French (0.0132‰), Dutch (0.0191‰), and Russian (0.0243‰). When the pronoun marks explicit definiteness, the finding may reflect *interference*, as noted above in connection with CCONJ PRON ADJ.

kMid	trigram	SKA	CEALfi	KKAmulti	KKAen
20	PUNCT PUNCT ADV	0.00023236	0.00069445	0.00072253	0.0002972

The last trigram in the top 20 is PUNCT PUNCT ADV, which seems to be more frequent in translations than in native texts. There are three exception source languages to this tendency, namely Norwegian (0.0176‰), Dutch (0.0638‰), and Spanish (0.1204‰), all of which were also outliers in the previous trigram PUNCT PUNCT PUNCT. As with the PUNCT PUNCT PUNCT and PUNCT PUNCT PROPN trigrams, the vast majority of individual occurrences contain a quotation mark or a quotation dash. A typical example is "*, "miten*" (', "how'). Thus, I suggest that the underlying phenomenon is also the same: more frequent use of quotation dashes in native texts than in translations, in which case the ends of quotes marked with quotation dashes do not come up here and raise the frequency. As the PUNCT PUNCT combination has come up three times in this top 20 list, I want to look at the bigram PUNCT PUNCT.

kMid	bigram	SKA	CEALfi	KKAmulti	KKAen
38	PUNCT PUNCT	0.00537676	0.03064121	0.01412816	0.02167528

The same tendency remains in the bigram. The gram is more frequent in translations and the individual occurrences virtually always contain a quotation mark or a quotation dash. In the KKAmulti figures differentiated by *origin*, there are two source languages whose figures fall below the native figure: Norwegian (1.243‰) and Spanish (4.521‰). Both of those source languages were also outliers in all the three trigrams containing the sequence PUNCT PUNCT above.

5.4 3+1-grams

The 3+1-grams are tetragrams that are sentence-terminal, i.e. they have the *end* value of 2. Thus, the fourth member is virtually always PUNCT. Cases where the *end* value is 3, i.e. the gram forms a sentence in itself, were not included in this set.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
1	PROPN VERB VERB PUNCT	0.00118983	0.00043549	0.0004368	0.0007519
2	SCONJ PROPN NOUN PUNCT	0.00037517	0.00021775	0.0002184	0.0002968
3	VERB PRON ADJ PUNCT	0.00170434	0.00221375	0.00220969	0.0021073
4	SCONJ NOUN VERB PUNCT	0.00267979	0.00152422	0.00151595	0.00180061
5	CCONJ PRON PUNCT PUNCT	7.5034E-05	0.00032662	0.00032118	0.0004551
6	PRON AUX NOUN PUNCT	0.00457708	0.00352023	0.00353293	0.00476864
7	AUX SCONJ NOUN PUNCT	0.00026798	0.00021775	0.0002184	0.00020776
8	VERB AUX ADJ PUNCT	0.00030014	0.00047178	0.00047534	0.00037595
9	ADP PUNCT ADV PUNCT	1.0719E-05	0.00010887	0.00010278	3.9574E-05
10	PRON ADJ NOUN PUNCT	0.00874682	0.0109599	0.01086859	0.00805327
11	INTJ PUNCT VERB PUNCT	0.00017151	3.6291E-05	3.8541E-05	7.9148E-05
12	ADP NOUN PUNCT PUNCT	0.00012863	0.00054437	0.00051388	0.00056393
13	AUX ADP NOUN PUNCT	0.00015007	0.00029033	0.00028263	0.00018798
14	NUM PRON NOUN PUNCT	5.3596E-05	0.00021775	0.00023125	0.00016819
15	PROPN VERB ADJ PUNCT	0.00026798	0.00014516	0.00014132	0.00023744
16	AUX AUX NOUN PUNCT	0.00199376	0.00239521	0.00241524	0.00292846
17	PRON ADJ ADJ PUNCT	0.00020366	0.00036291	0.00037256	0.0002968
18	ADV SCONJ NOUN PUNCT	0.001565	0.00097986	0.00100207	0.00173135
19	PRON ADV ADJ PUNCT	0.00077178	0.00141535	0.00137463	0.00104871
20	VERB PRON ADJ NOUN	1.0719E-05	3.6291E-05	3.8541E-05	2.968E-05

Table 6: The 20 most interesting 3+1-grams

Again, I will only go through the grams that yield meaningful results.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
1	PROPN VERB VERB PUNCT	0.00118983	0.00043549	0.0004368	0.0007519

The first 3+1-gram is PROPN VERB VERB PUNCT. It seems to be more frequent in native Finnish than in translations. There is one exception, namely Finnish translated from Swedish (2.128‰). The individual occurrences of the gram correspond to a structure where the proper noun acts as a subject for the following verb. The second verb is either an adverbial infinitival or a main verb, in which case the first verb is a modal auxiliary that has been tagged as a plain verb. A typical example of the structure is "*Martikainen halusi lähettää.*" ('Martikainen wanted to send.') The finding may reflect the same phenomenon already observed in connection with the VERB PUNCT bigram: clauses tend to end in verbs more often in native than in translated Finnish. Even the outlier source language matches (and

the other previous outlier, Norwegian, has the second highest relative frequency). However, in this gram's case, the phenomenon of native clauses ending in verbs more often than translated clauses does not seem to be a manifestation of the *conventionality* universal, at least in the sense of translations being less colloquial, as the structures are not markedly colloquial since the word order is unmarked (cf. VISK §1390), although it is possible that the subject (PROPN) is preceded by another argument of the main verb, rendering the word order marked after all.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
2	SCONJ PROPN NOUN PUNCT	0.00037517	0.00021775	0.0002184	0.0002968

The second 3+1-gram is SCONJ PROPN NOUN PUNCT. It also seems to be more frequent in native Finnish. There are, however, two source languages whose translations have a higher frequency: German (0.3756‰) and Hungarian (0.4834‰). In the individual occurrences, the subordinating conjunction is virtually always *kuin* ('as', 'like') and the following noun phrase is most often either a simile or a concrete comparison where the proper noun is in genitive. Some typical examples are "*kuin Topin isä.*" ('like Topi's father.'), "*kuin Asserin elin.*" ('like Asser's organ.'), and "*kuin Pohjanmaan lakeudet.*" ('like the plains in Ostrobothnia.').

We do not know yet whether similes are less frequent in translations *per se* or just less frequent in sentence-terminal positions. This could be controlled by somehow extracting all cases of similes from the corpora. However, that is beyond the scope of this study, because there is no clear correlation between any single non-position-disambiguated part-of-speech n-gram and similes.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
4	SCONJ NOUN VERB PUNCT	0.00267979	0.00152422	0.00151595	0.00180061

The next gram is SCONJ NOUN VERB PUNCT, which corresponds to a subordinate clause with two words. The noun is most often the subject of the clause, but, sometimes, it is an adverbial. The verb is the predicate of the clause. A typical example is "*kun mummo sairastui.*" ('when grandma fell sick."). The structure seems to be more frequent in native Finnish than in translated Finnish. There is only one exception in the KKAmulti figures differentiated by *origin*: Norwegian as a source language (2.937‰). Again, this finding may reflect the general tendency of native

clauses ending in verbs, but, this time, the correlation between the part-of-speech n-gram and the syntactical structure is especially strong.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
5	CCONJ PRON PUNCT PUNCT	7.5034E-05	0.00032662	0.00032118	0.0004551

The fifth gram in the list is CCONJ PRON PUNCT PUNCT. It seems to be more frequent in translations. Even all the KKAmulti figures differentiated by *origin* are in line this time. We see that the gram contains the sequence PUNCT PUNCT, which we have already gone through. Indeed, virtually all the individual occurrences contain a (closing) quotation mark. Roughly half of the occurrences are "*vai mitä?"*" ('right?', 'eh?', 'isn't it?'). The finding may reflect the already observed tendency of direct quotes being marked differently in native texts, but I also suggest that this finding is a manifestation of *interference*: in Finnish, the possible but not so common question tag is over-represented in translations because the question tag is more common in the source systems, and if there is a question tag in a source text, the translator might just leave it be, i.e. translate it explicitly.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
7	AUX SCONJ NOUN PUNCT	0.00026798	0.00021775	0.0002184	0.00020776

The seventh 3+1-gram in the list is AUX SCONJ NOUN PUNCT. It seems to be more frequent in native texts than in translations. However, there are four exceptions in the KKAmulti figures differentiated by *origin*: Dutch (0.3007‰), German (0.3130‰), Russian (0.4332‰), and Swedish (0.5804‰). There are, however, outliers to the other direction as well: Estonian, Spanish, Hungarian, and Norwegian show zero absolute frequency. The gram itself corresponds quite well to a simple simile construction *olla kuin x* ('is/are like x'), e.g. "*oli kuin kuiskaus.*" ('was like a whisper.'). Virtually all occurrences follow the pattern.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
8	VERB AUX ADJ PUNCT	0.00030014	0.00047178	0.00047534	0.00037595

The gram VERB AUX ADJ PUNCT seems to be more frequent in translations than in native texts. There are, however, two exception source languages: Swedish (0.1935‰) and Estonian (0.1240‰). The gram itself seems to favor a construction where there is a modal auxiliary verb and a copula followed by the predicative and

then the clause-terminal punctuation mark, for example "*osasi olla inhottava.*", "*saattoivat olla vaarallisia.*", and "*täytyykin olla kunnollisia*.". The modal auxiliaries seem to be quite randomly tagged either as a VERB or as an AUX. The only exception is *olla* ('be'), which is rendered consistently as an AUX (in this gram, quite ironically, as it is the main verb). This ambiguity might be due to the fact that all the Finnish verbs that Universal Dependencies regards as auxiliaries can also function as prototypical main verbs and have complete inflection paradigms, unlike, say, English auxiliaries.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
11	INTJ PUNCT VERB PUNCT	0.00017151	3.6291E-05	3.8541E-05	7.9148E-05

The eleventh gram is INTJ PUNCT VERB PUNCT. It seems to be more frequent in native texts. When I looked at the individual occurrences, I saw there are only five occurrences that are not tagging errors in total in all the translations. The native occurrences include a couple of expletives. The occurrences that are not expletives are of the form discourse particle + comma + verb + period/exclamation mark. A typical example would be "*Hei, pysähdytään.*" ('Hey, let's stop.'). All the occurrences are markedly colloquial, so the finding seems to support the *conventionality* universal where translations tend to avoid colloquial language.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
12	ADP NOUN PUNCT PUNCT	0.00012863	0.00054437	0.00051388	0.00056393

This gram, ADP NOUN PUNCT PUNCT, includes the sequence PUNCT PUNCT, which tends to mean direct quotes. Alas, virtually all of the occurrences do contain a quotation mark. As previously observed, direct quotes tend to be marked differently in native texts, and the same tendency continues here. The adposition and noun in the beginning of the gram do not often belong to the same phrase as most of the adpositions are postpositions. The KKAmulti figures differentiated by *origin* show that there are zero occurrences in the translations from Norwegian and Spanish, the two source languages that have been constant outliers in the direct quote grams. The relative frequency of the gram in Finnish translated from Dutch (0.00752‰) falls below the native frequency, and the relative frequency in Finnish translated from Swedish (0.1935‰) is only a hair above the native one.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
13	AUX ADP NOUN PUNCT	0.00015007	0.00029033	0.00028263	0.00018798

The gram AUX ADP NOUN PUNCT seems to be more frequent in translated language than native language, although there are four exceptions in the KKAmulti figures differentiated by *origin*: Estonian (0.1240‰), Dutch (0.0075‰), Spanish (0), and French (0). The structure behind the gram is copula + *pre*position + noun + punctuation mark, e.g. "*oli vailla mieltä!*" ('was insane!', literally 'was without a mind!') or "*olivat ilman kattoa.*" ('were without a roof.'). Even though most adpositions are natural postpositions in Finnish, and, thus, it would be easy to think that the finding may be a manifestation of source language *interference*, the actual adpositions here (mostly *vailla* and *ilman*, both meaning 'without') are ones that are either typically prepositions (*vailla*) or virtually always prepositions (*ilman*). However, there is a typical Finnish way of constructing the notion of lack, namely the abessive case. As such construction does not occur in the source languages apart from Estonian and Hungarian, the prepositional way could be over-represented as the abessive does not suggest itself as an automatic equivalent. This would be a typical case of Tirkkonen-Condit's (2004: 177–178) *under-representation of unique items*.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
14	NUM PRON NOUN PUNCT	5.3596E-05	0.00021775	0.00023125	0.00016819

The gram NUM PRON NOUN PUNCT seems to be more frequent in translations than in native texts. The only exception is Finnish translated from French (0 occurrences). The gram seems to favor two constructions. The first is *yksi* ('one') + possessive personal pronoun/demonstrative pronoun + noun in the elative case + punctuation, e.g. "*yksi hänen tyttäristään!*" ('one of his/her daughters!') or "*yksi näistä paperipalloista.*" ('one of these paper balls.'). The other construction is *yksi* + ainoa ('only/single') + noun + punctuation, e.g. "*yksi ainoa purkki.*" ('one single can.') or "*yksi ainoa ajatus.*" ('one single thought.').

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
16	AUX AUX NOUN PUNCT	0.00199376	0.00239521	0.00241524	0.00292846

The gram AUX AUX NOUN PUNCT seems to be more frequent in translations than in native texts. There are, however, three exceptions to this tendency in the KKAmulti figures differentiated by *origin*: Hungarian (1.209‰), French (1.879‰),

and Swedish (1.935‰) yield lower frequencies as source languages. The gram corresponds quite consistently to a structure where the copula *olla* (‘be’) is preceded by a modal auxiliary (e.g. “*täytyy olla valhetta!*” [‘must be a lie!’]) or the negation word *ei*, which counts as a verb in Finnish (e.g. “*ei ollut daami.*” [‘was not a dame.’]). The following noun is either a predicative, an existential subject, or an adverbial. This finding mirrors the previous gram VERB AUX ADJ PUNCT which was also more frequent in translations and included a copula being modally modified. This time, however, the auxiliary is actually tagged as one.

kMid	3+1-gram	SKA	CEALfi	KKAmulti	KKAen
17	PRON ADJ ADJ PUNCT	0.00020366	0.00036291	0.00037256	0.0002968
19	PRON ADV ADJ PUNCT	0.00077178	0.00141535	0.00137463	0.00104871

These two grams seem to be more frequent in translations than in native texts and reflect the same underlying phenomenon. Many of the occurrences follow the structure where a phrase that would be an adjective phrase by itself is turned into a noun phrase by the preceding pronoun. In the gram PRON ADJ ADJ PUNCT, the head word of the phrase is modified by another adjective, which is in the genitive case, e.g. “*jotakin sietämättömän tylsää.*” (‘something unbearably boring.’) or “*jotakin äärettömän tärkeätä.*” (‘something tremendously important.’). In the gram PRON ADV ADJ PUNCT, the head word of the phrase is modified by an adverb, e.g. “*jotakin täysin uutta.*” (‘something completely new.’) or “*jotakin erittäin tärkeää.*” (‘something very important.’). In the former gram, there are four outlier source languages: Swedish (0.1935‰), Dutch (0.1504‰), Norwegian (0), and Spanish (0). In the latter gram, there is only one source language below the native figure, namely Russian (0.7582‰). Despite these outliers, we might conclude that this phenomenon is a manifestation of *interference*, as the act of making an adjective phrase into a noun phrase with a pronoun is more marked in Finnish than in, say, English, because Finnish does not need additional morphemes to use an adjective (phrase) as a noun (phrase). In Finnish, we can say *nähdä uutta* (see new-PART, ‘see [something] new’) just as well as *nähdä ihmisiä* (see person-PART-PL ‘see people’) (cf. Bisang 2002: 2 about Quechua).

5.5 Conclusions about the method

In the vein of Borin and Prütz (2001), the method used in this study was to calculate relative frequencies part-of-speech n-grams in the different corpora and to harvest interesting n-grams by seeking out grams that are at the same time similar in frequency between translation corpora and different in frequency between native and translated text. The method of deciding difference/similarity was to calculate logarithms of ratios instead of Borin and Prütz' rank number difference.

The method proved to be somewhat usable. Although there was quite a lot of noise, the difference test (Figure 4) did pick out instances of clear differences between translations and native texts. A useful methodological finding was the fact that n-grams that are longer than a phrase disambiguate the part-of-speech string to better correspond to a specific syntactical structure, especially when the member of the gram that does not belong to the phrase anchors the gram to a clause border (e.g. PUNCT, SCONJ, and CCONJ). Many times, a phenomenon was found in the trigrams or 3-1-grams list even though the structure itself is two words long. In the future when conducting research like this, it is advisable to seek out the scope of the structure by looking at n+k-grams around the structure (giving *k* values —1, 1, 2 etc.). Perhaps, it would even be possible to automate such procedures.

The tagger was found to be not fully reliable in segmenting the data into sentences (see the trigram PUNCT PUNCT PUNCT). In the future, it would be advisable to segment the data with a stand-alone segmenter before other annotation is carried out.

5.6 Syntactical n-grams and other future possibilities

Even though we have looked at part-of-speech n-grams in this study, in the end, we are more interested in syntactical structures. In the future (and, perhaps, even in the present in different language pairs), when automatic parsing of natural language is more accurate than now, we might be able to drop the proxy of part-of-speech n-grams and study the dependency relations directly, perhaps even in a way where the unit of examination is not a word but a phrase, e.g. a trigram of noun phrase functioning as a subject, verb phrase as a predicate, and another noun phrase as object (cf. Sidorov et al. 2012: 1–5).

As a little test about how far from that situation we are, I looked more closely to the tendency of native clauses to end in verbs. The most common word order in the cases where the clause ending in a verb is marked (and markedly colloquial) is when the verb is preceded by either a direct object, e.g. "*Minä en unia näe.*" (literally 'I don't dreams see.') or an oblique nominal (see Universal Dependencies, f) (including cases that would count as an indirect object in English), e.g. "*Auvo ei minulle kertonut,*" (literally 'Auvo didn't [to] me tell,'). Thereby, I calculated all 5-grams that end in PUNCT, as the clauses need to be at least four words long, and picked out all the grams whose fourth member was simultaneously POS-tagged as VERB and syntactically analyzed as root (i.e. main verb). Of those grams that remained, I picked out all whose third member was syntactically analyzed as being either obl (oblique nominal) or obj (direct object). I picked out the first 20 instances of those grams in the corpora and analyzed the results, which are in Table 7 below.

corpus	total amount of grams	marked instances in sample	erroneously tagged n-grams in sample	estimated relative frequency[6]
SKA	1317	14	2	6.7064E-10
CEALfi	284	14	6	6.4260E-9
KKAmulti[7]	810	13	7	9.9568E-10
KKAen	829	10	8	6.2908E-10

Table 7: Sample test about marked verb-final word order

From the data in the table, we deduce that the accuracy of the dependency relation tagger is about 71.25%, which is quite low, which is in turn why I did not use the dependency relations in the main study (in comparison, the average accuracy of the POS tagger is over 90% [Charles University 2018]). However, despite the

[6] [(marked instances in sample/20)/total amount of grams]/total amount of 5-grams in corpus
[7] Due to the way the workflow was structured, all the samples are, unfortunately, from KKAru.

inaccuracy, the method of using dependency relations in such confined n-grams manages to pick out the relevant structures in over half of the cases.

When the estimated relative frequencies of the n-grams of marked verb-final word order are calculated, the previously extracted tendency disappears, as the CEALfi and KKAmulti corpora receive higher frequencies than the native SKA. Even though the estimate is very crude, it seems there is more to the verb-finality of native texts than the colloquial word order.

In addition to the phrasal dependency relation n-grams, a hybrid method could be further developed. It might prove useful to build corpus software (tagger and data processing tools) specifically for studying the tendencies of translated language. A part-of-speech tagging is a good starting point, but it should be possible to flexibly adjust the grain size of the category division. In this study, for example, it would have been useful if the PUNCT category would have been divided further into e.g. QUOTE, FULLSTOP, COMMA, and so forth, but also if the categories of AUX and VERB as well as PROPN and NOUN would have been combined. After this part-of-speech base layer, conditional dependency relation and morphological layers could be added. It would be good if one could search for, e.g. the bigram NOUN-nominative NOUN-adessive or the trigram NOUN-obj VERB-root PUNCT.

If the tendencies of translation could be reliably identified, we could teach machines to recognize, for example, typical syntactical interference patterns and avoid them. This would be useful in making better machine translation engines and in developing tools for automatic translation quality evaluation. The same awareness could also be used in translation training. Thus, the description and the theory, having feeded each other, would then have contributed to the translation applications, just as in Holmes' (1972/2004: 184) and Toury's (1995: 15–19) model.

5.7 Possible sources of errors

In this sub-chapter, I shall go through the problems in this study that I think can be sources of error. Firstly, all the translation from English (the CEALfi corpus) are by the same translator, Kersti Juva. Some properties of the language may be Juva's idiolect. This problem is mitigated by adding a second corpus of translated English–Finnish literary prose that contains translations by multiple translators.

A second source of error may be that the parser is not perfect. Computers make errors when interpreting natural languages. Not every word is marked to belong to the correct part-of-speech category. Ideally, the computer would make the same errors in all the corpora and the errors would thus cancel each other out, but there is no automatic way of knowing whether this is the case or not. However, John Sinclair, a pioneering computational linguist, asserts that "[s]ometimes the software may just get it wrong, and as long as there is no regular pattern to the mistakes, they are unlikely to have a great effect on the results of analysis when the corpus is many millions of words in length" (Sinclair 1992: 395).

However, none of my corpora reach even 2 million words in length. The amount of data might still be too small to make any more than very tentative suggestions. Especially, the KKAmulti sub-corpora differentiated by *origin*, i.e. single source languages different from English, consist only of one or two books each, and the effect of a single author's or translator's style plays a notable role in the figures.

As the amount of data is still relatively small, the rarer n-grams only have a handful of absolute occurrences. Should a rare trigram or 3+1-gram increase, by chance, from zero to one in absolute frequency in one of the small KKAmulti sub-corpora, it might overtake the native SKA corpus in *relative* frequency. I would feel much more confident with a dataset ten times as large as the present one.

Another problem with the KKAmulti corpus is that, despite including the Finno-Ugric Estonian and Hungarian as source languages, it still mainly consists of Finnish translated from Indo-European languages and, thus, typological differences between the Finno-Ugric languages (which Finnish belongs to) and Indo-European languages might over-emphasize some phenomena. A control corpus of translations from multiple source languages would be better if it had a more even distribution of source languages of different language families.

6 Conclusions

In this thesis, I have compared the relative frequencies of part-of-speech n-grams between native and translated Finnish literary prose, concentrating on such n-grams that are simultaneously similar in frequency between the CEALfi corpus of Finnish

translated from English and the KKAmulti corpus of Finnish translated from multiple source languages and different in frequency between the CEALfi corpus and the SKA corpus of native Finnish. The two most consistent findings are that there are more pronouns, especially personal pronouns, in translations than in native texts and that there are more verbs, especially in clause-final positions, in native texts than in translations.

The pronominal findings could be said to support the *interference*, *under-representation of unique items*, and (*T-*)*explicitation* universal candidates, as the more implicit and unique way of leaving a pronoun out does not occur in translations as often as in native texts, because the more explicit way of the source systems is transferred over the language boundary. The verbal findings do not clearly support any single universal candidates. More research is needed into the possibility of more marked, verb-final colloquial clauses in native texts.

Another phenomenon having to do with pronouns in translations is the tendency of translations having demonstrative pronouns denoting definiteness more often than native texts. This could also be said to support the same three universals: *interference*, *under-representation of unique items*, and (*T-*)*explicitation*. The notion of definiteness might be denoted more explicitly in translations, because the explicit way of marking it is transferred over to the translations and the uniquely Finnish, implicit way is under-represented.

A third manifestation of *interference* is the over-representation of question tags in translated language. Question tags are possible but not very common in native Finnish, but the source texts may trigger them into existence in the translations.

A general hypothesis could be made about these findings: whenever the target system has two or more ways of denoting something where the source system only has one, the way of the source system is over-represented and the other ways under-represented in the translations. This is *interference* as well as *under-representation of unique items of the target language*. In fact, I would claim the two are most often the same thing, unless the *interference* results in unacceptable translations.

There is also one phenomenon that could be said to support the *conventionality* universal: there are more interjections, especially expletives, in native texts than in

translations. As interjections are colloquial in nature, they are perhaps avoided in translations that might strive towards conservative and conventional language use.

Finally, a curious difference in the ways of marking direct quotes between native texts and translations was found. In native texts, the method of marking direct quotes with a quotation dash seems to be more common than in translated texts, where the common practice is to use quotation marks.

The method proved to be usable, and it should be developed further in the future, perhaps by using dependency relation n-grams instead of part-of-speech ones or by using a hybrid method. In addition, the strive towards better and larger corpora is perennial. Although the dark mirror of n-gram frequencies into translation universals is much clearer now than in the turn of the millennium, our knowledge about translated language is far from full.

References

Primary sources

The Corpus of Translated Finnish. Electric corpus for studying translated Finnish. Compiled at the Department of International Communication for the project Käännössuomi ja kääntämisen universaalit, University of Joensuu 1997–. Accessed: 2018-03-18. Availability: The server of the Department of International Communication, University of Joensuu.

Juva, Kersti 2018. *Englantilaisen ja amerikkalaisen kirjallisuuden klassikoita Kersti Juvan suomentamina, englanti–suomi-rinnakkaiskorpus* [tekstikorpus]. Received directly from author.

Secondary sources

Auvinen, Mira 2005. Geneerinen *sinä* käännössuomessa ja alkuperäissuomessa – kvantitatiivinen vertailu. In A. Mauranen, J. H. Jantunen (eds.) *Käännössuomeksi: Tutkimuksia suomennosten kielestä.* Tampere: Tampere University Press:71–83.

Baker, Mona 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In M. Baker, G. Francis, E. Tognini-Bonelli (eds.) *Text and Technology: In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins Publishing Company:233–250.

Baker, Mona 1995. Corpora in Translation Studies: an Overview and Some Suggestions for Future Research. In *Target* (7:2):223–243.

Baker, Mona 1996. Corpus-based translation studies: The challenges that lie ahead. In H. Somers (ed.) *Terminology, LSP and Translation: Studies in language engineering in honour of Juan C. Sager*. Amsterdam/Philadelphia: John Benjamins Publishing Company:175–186.

Bisang, Walter 2002. *Typology 6: Parts of speech*. Handout. 7th Summer School of the German Linguistic Society. Available at http://www.phil-fak.uni-duesseldorf.de/summerschool2002/Bisang6.PDF. Accessed on 2018-04-18.

Borin, Lars and Prütz, Klas 2001. Through a glass darkly: Part of speech distribution in original and translated text. In W. Daelemans, K. Sima'an, J. Veenstra, J. Zavrel (eds.) *Computational Linguistics in the Netherlands 2000*:30–44.

Charles University 2018. UDPipe User's Manual | ÚFAL. http://ufal.mff.cuni.cz/udpipe/users-manual#universal_dependencies_20_models. Accessed on 2018-04-26.

Chesterman, Andrew 2004. Beyond the particular. In A. Mauranen, P. Kujamäki (eds.) *Translation universals: Do they exist?* Philadelphia: John Benjamins Publishing Company:33–49.

Chipster. Chipster. https://chipster.csc.fi. Accessed on 2018-03-22.

Eskola, Sari 2004. Untypical frequencies in translated language. In A. Mauranen, P. Kujamäki (eds.) *Translation universals: Do they exist?* Philadelphia: John Benjamins Publishing Company:83–99.

Eskola, Sari 2005. Lauserakenteiden käytön piirteitä suomennetussa kaunokirjallisuudessa. In A. Mauranen, J. H. Jantunen (eds.) *Käännössuomeksi: Tutkimuksia suomennosten kielestä.* Tampere: Tampere University Press:225–243.

Even-Zohar, Itamar 1979. Polysystem Theory. In *Poetics Today*, Vol. 1, No. 1/2, Special Issue: Literature, Interpretation, Communication (Autumn, 1979). Durham: Duke University Press:287–310.

Frawley, William 1984. Prolegomenon to a Theory of Translation. In W. Frawley (ed.) *Translation: Literary, Linguistic, and Philosophical Perspectives*. Newark: University of Delaware Press, London and Toronto: Associated University Presses:159–178.

Hudson, Richard 1994. About 37% of word-tokens are nouns. In *Language*. 1994; 70 (2):331–339. Available at http://www.jstor.org.libproxy.helsinki.fi/stable/415831.

Heikkinen, Vesa; Lehtinen, Outi, and Lounela Mikko 2001. Kuvia kirjoitetusta suomesta. In *Kielikello* 2001(3). Available at http://www.kielikello.fi/index.php?mid=2&pid=11&aid=1279.

Holmes, James S. 1972/2004. The Name and Nature of Translation Studies. In L. Venuti (ed.) 2004. *The Translation Studies Reader: Second Edition*. New York and London: Routledge:180–192.

Jantunen, Jarmo Harri 2004. Untypical patterns in translation. In A. Mauranen, P. Kujamäki (eds.) *Translation universals: Do they exist?* Philadelphia: John Benjamins Publishing Company:101–126.

Jääskeläinen, Anni 2015. Suomen äännesymboliikkaa imitatiivien kautta tarkasteltuna. In *Virittäjä* 4/2015:464–497.

Juva, Kersti 2017. *Classics of English and American Literature in Finnish, Sentences and Paragraphs in the Original Order* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017011302

Juva, Kersti; Hartikainen, Kaija (ed.) 2014. Kersti Juva - 375 Humanistia. http://375humanistia.helsinki.fi/humanistit/kersti-juva-1. Accessed on 2018-04-23.

Kielipankki, a. Myllyn käyttöohjeet | Kielipankki. https://www.kielipankki.fi/tuki/mylly/. Accessed on 2018-02-05.

Kielipankki, b. Aineistot | Kielipankki. https://www.kielipankki.fi/aineistot/. Accessed on 2018-03-17.

Laury, Ritva 1996. *Sen* kategoriasta – onko suomessa jo artikkeli? In *Virittäjä* 2/1996:162–181.

Laviosa-Braithwaite, Sara 1996. *The English Comparable Corpus (ECC): A resource and a methodology for the empirical study of translation*. PhD thesis. Manchester: UMIST.

Mauranen, Anna 2000. Strange Strings in Translated Language. In M. Olohan (ed.) *Intercultural Faultlines: Research Models in Translation Studies I: Textual and Cognitive Aspects*. Manchester, UK & Northampton MA: St. Jerome Publishing:119–142.

Mauranen, Anna 2004. Corpora, universals and interference. In A. Mauranen, P. Kujamäki (eds.) *Translation universals: Do they exist?* Philadelphia: John Benjamins Publishing Company:65–82.

Mauranen, Anna and Tiittula, Liisa 2005. MINÄ käännössuomessa ja supisuomessa. In A. Mauranen, J. H. Jantunen (eds.) *Käännössuomeksi: Tutkimuksia suomennosten kielestä.* Tampere: Tampere University Press:35–69.

Olohan, Maeve 2004. *Introducing Corpora in Translation Studies*. London and New York: Routledge.

Pulla, Johanna 2011. *Käännösuniversaalit ja temporaalirakenne talouskielessä*. Pro gradu thesis, University of Tampere. Available at https://tampub.uta.fi/bitstream/handle/10024/82687/gradu05190.pdf.

Puurtinen, Tiina 2005. Käännössuomen piirteitä lastenkirjallisuudessa. In A. Mauranen, J. H. Jantunen (eds.) *Käännössuomeksi: Tutkimuksia suomennosten kielestä.* Tampere: Tampere University Press:211–223.

Sidorov, Grigori; Velasquez, Francisco; Stamatatos, Efstathios; Gelbukh, Alexander; and Chanona-Hernández, Liliana 2013. Syntactic Dependency-based N-grams as Classification Features. In I. Batyrshin, M. Gonzáles Mendoza (eds.) *Advances in Computational Intelligence: 11th Mexican International Conference on Artificial Intelligence, MICAI 2012 San Luis Potosí, Mexico, October 27 – November 4, 2012 Revised Selected Papers, Part II*. Berlin Heidelberg: Springer-Verlag:1–11. Available at http://www.cic.ipn.mx/~sidorov/sn_grams_MICAI2012.pdf.

Sinclair, John M. 1992. The automatic analysis of corpora. In J. Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991*. Berlin/New York: Mouton de Gruyter:379–397.

Tirkkonen-Condit, Sonja 2004. Unique items – over- or under-represented? In A. Mauranen, P. Kujamäki (eds.) *Translation universals: Do they exist?* Philadelphia: John Benjamins Publishing Company:177–184.

Toury, Gideon 1980. *In Search of a Theory of Translation*. Jerusalem: Academic Press.

Toury, Gideon 1995. *Descriptive Translation Studies – and beyond*. Amsterdam/Philadelphia: John Benjamins.

Universal Dependencies, a. UD v2. http://universaldependencies.org/v2/index.html. Accessed on 2018-02-08.

Universal Dependencies, b. Universal Dependencies. http://universaldependencies.org. Accessed on 2018-03-27.

Universal Dependencies, c. Universal POS tags. http://universaldependencies.org/u/pos/index.html. Accessed on 2018-03-27.

Universal Dependencies, d. AUX. http://universaldependencies.org/fi/pos/AUX_.html. Accessed on 2018-03-31.

Universal Dependencies, e. X. http://universaldependencies.org/u/pos/X.html. Accessed on 2018-04-03.

Universal Dependencies, f. obl. http://universaldependencies.org/u/dep/obl.html. Accessed on 2018-04-15.

VISK = Hakulinen, Auli; Vilkuna, Maria; Korhonen, Riitta; Koivisto, Vesa; Heinonen, Tarja Riitta; and Alho, Irja 2004. *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura. Online version, accessed 2018-04-06. Available at http://scripta.kotus.fi/visk. URN:ISBN:978-952-5446-35-7.

SUOMENKIELINEN LYHENNELMÄ

Helsingin yliopisto
Käännös- ja tulkkausviestintä, englanti
Matias Tamminen: Kerran tietoni on täydellistä: sanaluokka-n-grammien suhteelliset frekvenssit käännetyssä ja supisuomalaisessa kaunokirjallisessa proosassa
Pro gradu -tutkielma 58 s., suomenkielinen lyhennelmä 12 s.
toukokuu 2018

---

1 Johdanto

Tutkin pro gradu -tutkielmassani sanaluokka-n-grammien eli sanaluokkatagattujen sanojen muodostamien, *n* sanaa pitkien ketjujen (esim. 3-grammi NOUN VERB SCONJ) suhteellisten frekvenssien eroja käännetyssä ja supisuomalaisessa kaunokirjallisessa proosassa. Pyrin pääsemään käsiksi käännetylle suomelle ominaisiin syntaktisiin piirteisiin. Haluan selvittää, ovatko mahdolliset erot supisuomen ja käännösten välillä odotuksenmukaisia eri käännösuniversaalihypoteesien eli käännöksille ominaisten, systemaattisten, kieliparista ja -suunnasta riippumattomien taipumusten (kts. esim. Baker 1993: 242–245) valossa.

Tämänkaltaiselle tutkimukselle on tarvetta, koska tutkimuksesta saadun tiedon avulla voidaan esimerkiksi oppia, minkälaisissa rakenteissa tyypillisesti esiintyy lähtötekstin siirrännäisvaikutusta eli *interferenssiä*. Tätä tietoa voidaan sitten käyttää muun muassa koneoppimisessa opettamaan koneita tunnistamaan tällaiset interferenssin ilmentymät ja kääntäjänkoulutuksessa osoittamaan tuleville kääntäjille tyypillisiä interferenssirakenteita, jotta he voivat tunnistaa ne ja välttää niitä omissa käännöksissään.

Esittelen ensiksi luvussa kaksi deskriptiivisen käännöstieteen kentän ja korpuspohjaisen käännöstutkimuksen historian. Sen jälkeen esittelen käyttämäni korpukset ja metodin luvussa kolme. Luvussa neljä esittelen ja analysoin tulokseni. Lopuksi, luvussa viisi, esitän johtopäätökseni.

2 Teoreettinen tausta

Esittelen tässä luvussa deskriptiivisen käännöstutkimuksen kentän, käännösuniversaalien konseptin ja korpuspohjaisen käännöstutkimuksen historian.

2.1 Lähtökielestä ja preskriptiosta kohdekieleen ja deskriptioon

Käännöstieteessä tutkimuksen keskiössä on historiallisesti ollut lähtöteksti. Käännöksiä on tutkittu vertaamalla niitä lähtötekseihin ja käännösratkaisuja arvioitu sen perusteella, miten ekvivalentteja ne ovat lähtötekstin ratkaisuihin verrattuna. (Baker 1993: 233–235.) Tämä asetelma alkoi muuttua 1990-luvulla, kun tutkimuskentällä esitettiin toiveita fokuksen siirtämiseksi lähtö- ja kohdetekstin välisestä suhteesta kohdetekstin ja kohdesysteemin väliseen suhteeseen (Baker 1993: 236) ja preskriptiosta (miten asioiden tulisi olla) deskriptioon (miten asiat todellisuudessa ovat) (Toury 1995: 1–5).

Siirtyminen preskriptiosta deskriptioon johtui pitkälti kasvavasta metodologisesta tyytymättömyydestä. Käännöstieteestä haluttiin metodologisesti systemaattinen. (Toury 1980: 81, Baker 1993: 240.) Ensimmäisenä idean esitti James Holmes, joka sisällytti käännöstiedettä jaottelevaan karttaansa alaluokan deskriptiivinen käännöstiede (’Descriptive Translation Studies’). Deskriptiivinen käännöstiede muodosti yhdessä teoreettisen käännöstieteen kanssa puhtaan (’pure’) käännöstieteen, jonka Holmes rinnasti soveltavaan käännöstieteeseen. (Toury 1995: 9–10, Holmes 1972/2004: 184.) Deskriptiivisen käännöstieteen ideana on tutkia havaittavissa olevia, teoreettisesta viitekehyksestä riippumattomia faktoja ja testata käännösteorioiden tuottamia hypoteeseja (Toury 1980: 80). Toury jalosti Holmesin karttaa. Hänen mukaansa deskriptiivisten käännöstutkimusten empiirisistä havainnoista tulisi yleistää yleismaailmallisia teorioita, joita sitten jälleen koeteltaisiin deskriptiivisin menetelmin. Täten deskriptiivinen käännöstiede ja teoreettinen käännöstiede ikään kuin syöttäisivät toisiaan. Soveltavan käännöstieteen harjoittajat, esimerkiksi kääntämisen opettajat ja käännöskriitikot, voisivat sitten tehdä omat johtopäätöksensä siitä, minkälainen on hyvä käännös, mutta puhtaan ja soveltavan käännöstieteen välisen suhteen tulisi Touryn mukaan olla yksisuuntainen ja epäsuora. (Laviosa-Braithwaite 1996: 24–25, Toury 1995: 15–19.)

Preskriptiosta deskriptioon siirtyminen lomittui sen kanssa, että tutkimuksellinen kiinnostus siirtyi lähtötekstistä kohdetekstiin. Kohdetekstiorientaatio rakentui Itamar Even-Zoharin polysysteemiteorialle (Baker 1993: 237–238). Even-Zoharin mukaan standardia kielivarianttia ei voi tutkia ilman ei-standardin tutkimista, eikä käännetty kirjallisuus ole irrallaan kohdesysteemin alkuperäiskirjallisuudesta (Even-Zohar 1979: 292). Tämä ajattelutapa validoi tutkimussuunnan, jossa tutkitaan käännettyä kieltä ja sen suhdetta vastaavaan alkuperäiskieleen (Baker: 1993: 242). Kohdetekstiorientaatiota vahvisti myös Frawley, jonka mukaan käännökset muodostavat niin sanotun kolmannen koodin (’third code’), systeemin, joka eroaa niin lähdesysteemistä kuin kohdesysteemistäkin (Frawley 1984: 168–169).

2.2 Käännösuniversaalit

Kohdetekstiorientaatioon siirtyminen johti käännösuniversaaliajatuksen syntyyn. Käännösuniversaalit ovat kaikelle käännetylle tekstille yhteisiä taipumuksia (Baker 1993: 242). Baker listaa ehdotuksikseen tällaisiksi universaaleiksi *eksplikaation* (käännökset ovat eksplisiittisempiä kuin lähtötekstit ja vastaavat alkuperäistekstit kohdesysteemissä), *simplifikaation/disambiguaation* (käännökset ovat syntaktisesti yksinkertaisempia ja yksiselitteisempiä kuin niiden lähtötekstit), *konventionaalisuuden* (epätyypilliset tai kielenvastaiset elementit vaihtuvat tyypillisiksi elementeiksi käännöksissä), *toiston välttämisen*, *kohdekielen ominaispiirteiden ylikorostumisen* ja *epätyypilliset frekvenssit* (jotkin piirteet ovat yleisempiä tai harvinaisempia käännöksissä kuin vastaavissa kohdesysteemin kotoperäisissä teksteissä) (Baker 1993: 243–245).

Bakerin luetteloa on täydennetty myöhemmin. Uusia lisäyksiä ovat *kohdekielen uniikkiainesten aliedustuminen* (Tirkkonen-Condit 2004: 177–178), *epätyypilliset kollokaatiot* (kohdekielen vastaaviin kotoperäisiin teksteihin verrattuna) (Mauranen 200: 120) ja *interferenssi* (lähtöteksti/systeemi vaikuttaa käännöksiin) (Toury 1995: 274–275).

Kuten ehdotettujen universaalien luettelosta nähdään, osa universaaleista vertaa käännöksiä lähtöteksteihin ja osa vastaaviin kohdesysteemin alkuperäisteksteihin. Ensimmäinen eksplisiittisesti tästä erosta kirjoittanut henkilö oli Andrew Chesterman (2004: 39–40), joka kutsuu ensimmäistä luokkaa *S-universaaleiksi* ja jälkimmäistä *T-*

universaaleiksi. Keskityn tässä tutkielmassa lähinnä *T-universaaleihin* (ja *interferenssiin*, josta voidaan esittää typologisia arvauksia), koska Borinilta ja Prützilta (2001) lainaamani ja edelleen kehittämäni metodi on suunniteltu kohdetekstien ja verrannollisten alkuperäistekstien vertaamista varten. Siksi käyttämäni rinnakkaiskorpuksen lähtötekstipuoli (CEALen, katso luku 3.3.1) jää valitettavan vähälle käytölle.

2.3 Korpukset deskriptiivisen käännöstieteen työkaluina

Deskriptiivisen käännöstieteen tarjotessa empiirisen tavan ajatella ja käännösuniversaalien tarjotessa testattavat hypoteesit tarvitaan enää aineisto ja metodi. Käännösuniversaaleja tutkittaessa tarvitaan korpuksia eli tekstikokoelmia, joiden tekstit on valittu tietyin kriteerein ja jotka ovat sähköisessä muodossa (Olohan 2004: 1).

Baker esittää toiveen, että korpuksia käytettäisiin tutkimusvälineinä tutkimuksissa, joissa pyritään selvittämään käännösten ja vastaavien kohdesysteemin alkuperäistekstien välisiä eroja. Näin voidaan testata esitettyjä käännösuniversaalihypoteeseja ja kenties esittää myös uusia universaaliehdokkaita. (Baker 1993: 245.) Baker (1995: 230–235) tarjoaa myös käännöstieteellisten korpusten typologian:

- *Rinnakkaiskorpukset* sisältävät käännöksiä ja niiden lähtötekstejä toisiinsa kohdistettuina.
- *Monikieliset korpukset* sisältävät useita perinteisiä yksikielisiä korpuksia useilla eri kielillä.
- *Verrannolliset korpukset* sisältävät käännöksiä ja samalla kielellä alun perin kirjoitettuja, samaa genreä edustavia verrannollisia tekstejä.

Käytän tutkimuksessani kahta korpusta: *Englantilaisen ja amerikkalaisen kirjallisuuden klassikoita Kersti Juvan suomentamina, englanti–suomi-rinnakkaiskorpusta* (Juva 2018) ja *Käännössuomen korpusta*, joka on lähes 20 vuoden iästään huolimatta suurin suomenkielinen verrannollinen korpus.

3 Materiaali ja metodi

Esittelen tässä luvussa tutkimusmateriaalini ja -metodini. Käsittelen aluksi idean käyttää sanaluokka-n-grammeja syntaksin välikappaleena. Sen jälkeen esittelen käyttämäni korpukset ja lopuksi sen, miten poimin mielenkiintoiset tapaukset lähempään tarkasteluun.

3.1 Sanaluokka-n-grammit syntaktisten rakenteiden approksimaatioina

Olen kiinnostunut käännetyn kielen syntaktisista piirteistä. Suomi ja muut agglutinoivat kielet ovat kuitenkin hyvin vaikeita jäsentää automaattisesti, joten tarkastelen sanaluokka-n-grammeja syntaktisten rakenteiden sijaan, koska lauseenjäsenkategorioilla on usein jokin tietty sanaluokkakategoria, jota ne suosivat. Esimerkiksi subjekti on useimmiten substantiivi. Jos tutkisin jotain koneellisten parsereitten kannalta helpompaa kieliparia (kuten Borinin ja Prützin ruotsi–englanti-paria), tarkastelisin lauseenjäsen-n-grammeja.

N-grammi on perättäisten yksiköiden *n* yksikköä pitkä ketju. Sanaluokka-n-grammi on n-grammi, jossa yksiköt ovat sanoja ja jossa yksiköiden tarkasteltu piirre on niiden sanaluokka.

Here	is	an	example	.
ADV	VERB	DET	NOUN	PUNCT

Taulukko 8: N-grammiesimerkkivirke

Yllä olevassa esimerkkivirkkeessä on

- viisi unigrammia (ADV, VERB, DET, NOUN, PUNCT)
- neljä bigrammia (ADV VERB, VERB DET, DET NOUN, NOUN PUNCT)
- kolme trigrammia (ADV VERB DET, VERB DET NOUN, DET NOUN PUNCT)
- kaksi 4-grammia (ADV VERB DET NOUN, VERB DET NOUN PUNCT)
- yksi 5-grammi (ADV VERB DET NOUN PUNCT).

Kuten esimerkistä käy ilmi, välimerkit lasketaan sanoiksi ja välimerkki (PUNCT) on oma sanaluokkakategoriansa.

3.2 Korpukset

Käytän tutkimuksessani kahta korpusta, jotka jaan pienempiin osakorpuksiin. Esittelen korpukset ja niiden alajaot tässä alaluvussa.

3.2.1 Englantilaisen ja amerikkalaisen kirjallisuuden klassikoita Kersti Juvan suomentamina

Käännössuomen pääkorpukseni on *Englantilaisen ja amerikkalaisen kirjallisuuden klassikoita Kersti Juvan suomentamina, englanti–suomi-rinnakkaiskorpus* (jatkossa CEAL) (Juva 2018). Se sisältää Kersti Juvan suomennokset Jane Austenin romaanista *Pride and Prejudice*, Henry Jamesin romaanista *Washington Square* ja Charles Dickensin romaanista *Bleak House* sekä näiden englanninkieliset alkutekstit kappaletasolla kohdistettuina (Juva 2018). Korpuksen suomenkielinen osa (jatkossa CEALfi) sisältää 502 062 sanaa ja englanninkielinen osa (jatkossa CEALen) 657 986 sanaa.

Valitsin tämän korpuksen, koska se on yksi harvoista englanti–suomi-rinnakkaiskorpuksista ja koska Kersti Juva on ehkä tunnetuin suomalainen kääntäjä, ja hänen käännöksensä ovat hyvin arvostettuja (Juva ja Hartikainen 2014).

3.2.2 Käännössuomen korpus

Toinen käyttämäni korpus on *Käännössuomen korpus* (jatkossa KSK). Korpus sisältää sekä supisuomalaisia että suomeksi monista eri lähtökielistä käännettyjä tekstejä monissa eri genreissä. KSK on siis verrannollinen korpus. Käytän KSK:sta ainoastaan supisuomalaisen kaunokirjallisen proosan osakorpusta (jatkossa SKA), englannista suomennetun kaunokirjallisen proosan osakorpusta (jatkossa KKAen), venäjästä suomennetun kaunokirjallisen proosan osakorpusta (jatkossa KKAru) ja monista eri kielistä (saksa, ranska, hollanti, norja, ruotsi, viro, unkari) suomennetun kaunokirjallisen proosan osakorpusta (jatkossa KKAmuut). SKA sisältää 1 212 770 sanaa ja KKAen 1 410 281 sanaa.

Poimin KKAru-korpuksesta ensimmäiset kaksi kirjaa ja yhdistin ne KKAmuut-korpukseen. Tuloksena on monilähtökielinen verrokkikorpukseni KKAmulti, jota käytän lähtökielen vaikutuksen kontrolloimiseen. En sisällyttänyt KKAru-korpusta kokonaan, jottei venäjä painottuisi kohtuuttomasti lähtökielenä. KKAmulti sisältää 1 148 215 sanaa.

3.2.3 Korpusten alajakojen yhteenveto

Käyttämäni korpukset alajakoineen on esitetty lyhyesti alla olevassa taulukossa.

lyhenne	käyttötarkoitus/sisältö	koko (sanaa)
CEALfi	englannista suomennetun suomen pääkorpus	502 062
SKA	supisuomen pääkorpus	1 212 770
KKAmulti	monilähtökielinen kontrollikorpus (sisältää KKAmuut-korpuksen ja kaksi kirjaa KKAru-korpuksesta)	1 148 215
KKAen	monien eri kääntäjien englannista suomentamien tekstien kontrollikorpus	1 410 281
CEALen	CEALfi-korpuksen tekstien lähtötekstit	657 986

Taulukko 9: Korpusten alajakojen yhteenveto

3.2.4 Korpusten preparointi

Käytän korpusdatan käsittelyssä Kielipankin Mylly-työkalua (Kielipankki, a).

Kaikki korpukset olivat aluksi pelkkää raakatekstiä, kukin kirja omana tiedostonaan. Ajoin Universal Dependencies 2 -parserin (Universal Dependencies, a) kaikille

kirjoille. Yhdistin sitten kirjat tiedostoihin osakorpuksittain siten, että kaikki saman osakorpuksen kirjat ovat samassa tiedostossa. Selvitin sitten näistä korpustiedostoista sanaluokka-n-grammit *n*:n arvoilla 1–4.

Kun n-grammit oli selvitetty, laskin kunkin n-grammin absoluuttisen frekvenssin kaikissa korpuksissa. Bi- ja trigrammien kohdalla laskin absoluuttiset frekvenssit kahdella tavalla: paljaat frekvenssit ja parametrin *end* mukaan erotellut frekvenssit. Parametri *end* kuvaa sitä, koskettaako grammi virkerajaa. Arvolla 0 grammi on virkkeen keskellä eikä kosketa virkerajaa. Arvolla 1 grammi on virkkeenalkuinen, arvolla 2 virkkeenloppuinen ja arvolla 3 kokonainen virke itsessään. Tetragrammien kohdalla laskin ainoastaan nämä *end*-parametrin mukaan erotellut arvot, sillä olen kiinnostunut ainoastaan 3+1-grammeista, joissa neljäs jäsen on virkkeenloppuinen lopetusmerkki. Tämän rajauksen tein siksi, että suurilla *n*:n arvoilla tulokset siroavat hyvin paljon.

Paljaiden frekvenssien ja *end*-eroteltujen frekvenssien lisäksi erottelin KKAmulti-osakorpuksen kohdalla frekvenssit *origin*-parametrin eli lähtökielen mukaan.

Tähän asti kaikki frekvenssit olivat olleet absoluuttisia, mutta tässä vaiheessa normalisoin frekvenssit eli muutin ne suhteellisiksi.

3.3 Korpustenvälisten erojen löytäminen

Kun n-grammien frekvenssit oli normalisoitu, laskin kunkin n-grammin kahden korpuksen välisen esiintyvyyden eron laskemalla suhteellisten frekvenssien osamäärän (kymmenkantaisen) logaritmin. Mitä suurempi logaritmin itseisarvo on, sitä merkittävämpi n-grammin esiintyvyyden ero korpusten välillä on. Logaritmin etumerkki kertoo eron suunnan.

3.4 Kontrollikorpuksen lisääminen yhtälöön

Tutkimukseni tarkoituksen on löytää käännössuomen yleisiä ominaispiirteitä. CEALfi- ja SKA-korpusten väliset erot voivat kuitenkin olla pelkästään englanti–suomi-kieliparille ominaisia. Siksi tarvitsen KKAmulti-korpusta, jossa on käännöksiä monista eri lähtökielistä. Olen siis kiinnostunut n-grammeista, joiden esiintyvyys on yhtä aikaa samankaltainen CEALfi- ja KKAmulti-korpuksissa ja erilainen CEALfi-

ja SKA-korpuksissa. Tällaiset grammit löydän alla olevalla lausekkeella.

$$lg\left(\frac{\left|lg\left(\frac{x}{y}\right)\right|}{\left|lg\left(\frac{x}{z}\right)\right|}\right)$$

Kuva 1: Erojen vertailun lauseke

Lausekkeessa *x* on grammin suhteellinen frekvenssi CEALfi-korpuksessa, *y* grammin suhteellinen frekvenssi SKA-korpuksessa ja *z* grammin suhteellinen frekvenssi KKAmulti-korpuksessa. Jos grammin frekvensseissä on suuri ero CEALfi- ja SKA-korpusten välillä, osoittaja saa suuren arvon. Jos grammin frekvensseissä on vain pieni ero CEALfi- ja KKAmulti-korpusten välillä, nimittäjä saa pienen arvon. Jos molemmat tapahtuvat yhtä aikaa, koko lausekkeen arvo on paljon nollan yläpuolella.

Järjestin n-grammit kullakin *n*:n arvolla yllä olevan lausekkeen arvon mukaiseen järjestykseen siten, että suurin arvo on ensimmäisenä. Keskityn tarkastelemaan näin saatujen grammilistojen kahdenkymmenen kärkeä.

4 Tulokset

Tässä luvussa käyn läpi saamiani tuloksia ja katson, voiko tulosten perusteella sanoa mitään esitetyistä käännösuniversaalihypoteeseistä.

4.1 Unigrammit

Kaikista systemaattisimmin erilainen sanaluokkakategoria supisuomen ja käännössuomen välillä on verbi. Verbien suhteellinen frekvenssi on 16,43 % SKA:ssa, 15,01 % CEALfi:ssä, 15,16 % KKAmultissa ja 15,69 % KKAen-korpuksessa, joten verbit vaikuttavat olevan yleisempiä supisuomessa kuin käännöksissä.

Toisin kuin verbit, pronominit vaikuttavat olevan yleisempiä käännössuomessa kuin supisuomessa (9,80 % SKA:ssa, 12,31 % CEALfi:ssä, 11,03 % KKAmultissa, 11,79 % KKAen-korpuksessa). Löydös on linjassa Maurasen ja Tiittulan (2005: 42) sekä Auvisen (2005: 77) havaintojen kanssa. Erityisesti persoonapronominit ovat

frekventimpiä käännöksissä kuin supisuomessa, minkä voi sanoa tukevan ainakin kolmea käännösuniversaalia: *interferenssiä*, *uniikkiainesten aliedustumista* ja (*T-*)*eksplikaatiota*. Interferenssi ja uniikkiainesten aliedustuminen ovat saman kolikon kaksi eri puolta. Suomessa substantiivia määrittävän genetiivisen persoonapronominin voi jättää ilmipanematta, koska omistussuhde näkyy possessiivisuffiksissa. Jos alkutekstissä persoonapronomini on ilmipantu (lähtösysteemin vaatimusten takia), se siirtyy helposti käännökseen, koska suomelle uniikki mahdollisuus jättää pronomini pois ei ehdota itseään vastineeksi. (T-)eksplikaatiota puolestaan on se, että käännöksissä omistussuhde näkyy kahdesti, sekä itse pronominissa että possessiivisuffiksissa, kun vastaavissa supisuomalaisissa teksteissä on useammin pelkkä suffiksi.

4.2 Muut grammit

Supisuomessa lauseet loppuvat useammin verbiin kuin käännössuomessa. Tämä ilmiö näkyy grammeissa VERB PUNCT, VERB CCONJ,
PROPN VERB VERB PUNCT ja SCONJ NOUN VERB PUNCT, jotka kaikki ovat yleisempiä supisuomessa kuin käännössuomessa. Verbiloppuinen sanajärjestys on kohosteinen reaktiivisella ja affektiivisella tavalla, mikäli verbi seuraa määreitään ja teemapaikkaa edeltävässä esikentässä on materiaalia (VISK § 1390). Tällaiset tapaukset ovat jokseenkin informaaleja ja voisivat käännöksissä aliedustuessaan tukea *konventionaalisuus*-universaalihypoteesiä (käännöksissä vältetään kirjakielen konventioiden vastaisia rakenteita), mutta teema–reema-analyysiä ei voida tällä hetkellä suorittaa koneellisesti, joten tarkempi tarkastelu jää tulevaisuuteen.

Supisuomessa vaikuttaa olevan enemmän kirosanoja kuin käännössuomessa. Tämä ilmiö näkyy grammeissa AUX INTJ, INTJ AUX, INTJ SCONJ ja
INTJ PUNCT VERB PUNCT, jotka ovat kaikki yleisempiä supisuomessa kuin käännössuomessa. Kaikki interjektiot eivät suinkaan ole kirosanoja, mutta yksittäisiä esiintymiä tarkasteltaessa käytännössä kaikki kirosanat ovat supisuomen aineistosta. Tämän löydöksen voi sanoa tukevan *konventionaalisuus*-universaalihypoteesiä, sillä kirosanojen voi sanoa olevan puhekielisiä ja informaaleja. Löydös on linjassa Puurtisen (2005) havaintojen kanssa. Puurtinen (2005) havaitsi, että lastenkirjojen käännöksissä on vähemmän puhekielisyyttä kuin vastaavissa supisuomalaisissa teksteissä.

Supisuomessa käytetään repliikkiviivaa suoran lainauksen merkkinä useammin kuin käännössuomessa. Tämä ilmiö käy ilmi grammeista PUNCT PUNCT PROPN, PUNCT PUNCT PUNCT, PUNCT PUNCT ADV, PUNCT PUNCT ja CCONJ PRON PUNCT PUNCT, jotka ovat kaikki yleisempiä käännössuomessa kuin supisuomessa. Jos suora lainaus merkitään lainausmerkein, sitaatin lopussa on useimmiten kaksi peräkkäistä välimerkkiä. Jos suora lainaus merkitään repliikkiviivalla, sitaatin lopussa on vain yksi välimerkki, jolloin grammit, joissa on kaksi peräkkäistä välimerkkiä, aliedustuvat.

Käännössuomessa vaikutetaan merkitsevän määräisyyttä eksplisiittisesti demonstratiivipronominillä useammin kuin supisuomessa. Tämä käy ilmi grammeista CCONJ PRON NUM, PRON NUM, CCONJ PRON ADJ ja PRON NUM PUNCT, jotka ovat kaikki yleisempiä käännössuomessa kuin supisuomessa. Monissa esiintymissä pronomini on nimenomaan demonstratiivipronomini, joka toimii määräisen artikkelin tapaan (vertaa Laury 1996). Löydöksen voi sanoa tukevan *interferenssi*-käännösuniversaalia, mikäli lähtötekstien eksplisiittinen määräisyys siirtyy käännöksiin.

5 Johtopäätökset

Olen pro gradu -tutkielmassani tarkastellut sanaluokka-n-grammien suhteellisten frekvenssien eroja käännetyssä ja supisuomalaisessa kaunokirjallisessa proosassa. Tulosten perusteella voisi yleistää, että mikäli lähtökielen systeemissä on vain yksi tapa ilmaista jokin tietty asia ja kohdekielen systeemissä lähtökielen systeemin tavan lisäksi muita vaihtoehtoisia tapoja, lähtökielen systeemin tapa yliedustuu käännöksissä ja muut tavat aliedustuvat. Tällaiset tapaukset ovat sekä *interferenssiä* että *kohdekielen uniikkiainesten aliedustumista*. Voisin jopa väittää, että nämä kaksi universaalia ovat useimmiten sama asia.

Interferenssin ja uniikkiainesten aliedustumisen lisäksi *konventionaalisuus*-käännösuniversaali saa tukea aineistostani. Se näkyy niin interjektioissa (eritoten kirosanoissa) että mahdollisesti verbiloppuisessa sanajärjestyksessä, tosin tätä sanajärjestyshavaintoa täytyy ehdottomasti tutkia enemmän ennen vahvempien johtopäätösten vetämistä.

Sanaluokka-n-grammien frekvenssien vertailu osoittautui kohtuullisen käyttökelpoiseksi metodiksi. Jatkossa metodia voisi kehittää käyttämällä sanaluokka-n-grammien sijaan esimerkiksi dependenssisuhteiden n-grammeja tai näiden kahden hybridiannotaatiota.